



# Multi-Type Outer Product-Based Fusion of Respiratory Sounds for Detecting COVID-19

Adria Mallol-Ragolta<sup>1</sup>, Helena Cuesta<sup>2</sup>, Emilia Gómez<sup>2,3</sup>, and Björn W. Schuller<sup>1,4,5</sup>

<sup>1</sup> EIHG – Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany

<sup>2</sup> MTG – Music Technology Group, Universitat Pompeu Fabra, Spain

<sup>3</sup> Joint Research Centre, European Commission, Spain

<sup>4</sup> Center for Interdisciplinary Health Research, University of Augsburg, Germany

<sup>5</sup> GLAM – Group on Language, Audio, & Music, Imperial College London, UK

adria.mallol-ragolta@informatik.uni-augsburg.de

## Abstract

This work presents an outer product-based approach to fuse the embedded representations learnt from the spectrograms of cough, breath, and speech samples for the automatic detection of COVID-19. To extract deep learnt representations from the spectrograms, we compare the performance of specific Convolutional Neural Networks (CNNs) trained from scratch and ResNet18-based CNNs fine-tuned for the task at hand. Furthermore, we investigate whether the patients' sex and the use of contextual attention mechanisms are beneficial. Our experiments use the dataset released as part of the Second Diagnosing COVID-19 using Acoustics (DiCOVA) Challenge. The results suggest the suitability of fusing breath and speech information to detect COVID-19. An Area Under the Curve (AUC) of 84.06 % is obtained on the test partition when using specific CNNs trained from scratch with contextual attention mechanisms. When using ResNet18-based CNNs for feature extraction, the baseline model scores the highest performance with an AUC of 84.26 %.

**Index Terms:** COVID-19 Detection, Respiratory Diagnosis, Transfer Learning, Information Fusion, Healthcare

## 1. Introduction

Digital health solutions powered with *Artificial Intelligence* (AI) have the potential to revolutionise the health care systems worldwide, improving the early diagnosis of diseases, and the monitoring of the patients towards personalised treatment plans. Previous works in the literature explored the use of AI-based techniques in a wide range of medical problems, including the detection of coughs or sneezes [1], or the recognition of mental illnesses, such as depression [2, 3] or *Post-Traumatic Stress Disorder* (PTSD) [4]. Such technologies do not aim at replacing medical diagnostic tools, rather providing scalable, cost-effective pre-screening solutions to optimise the medical resources.

Motivated by the outbreak of the *Coronavirus Disease 2019* (COVID-19) and its symptomatology, which presents affectations in the human respiratory system, researchers have investigated the use of AI techniques to detect the virus from the analysis of chest CT scans [5, 6] or X-ray images [7]. As the respiratory sounds might also be impacted because of the virus symptomatology, it seems reasonable to argue about their potential to contain salient information to detect COVID-19. In this regard, researchers have exploited cough [8, 9, 10], breath [11, 12], and speech [13, 14] information for this task.

The majority of previous works found in the literature present mono-type solutions for the automatic detection of COVID-19 patients. The largest body of knowledge exploits

cough-related information for this task [15, 16, 17, 18]. Other works also propose mono-type systems, but focusing only on the analysis of speech signals [19, 20]. The amount of works exploiting multi-type solutions to enrich and complement the salient information embedded in individual respiratory sounds is rather scarce. Despite the multi-type nature of the approach presented by Nessiem et al. [21], the authors model breathing and coughing sounds jointly, but only adding heterogeneity to the training data. Instead, Liu et al. [22] present a multi-type deep fusion method exploiting breathing and coughing information for the automatic detection of COVID-19 patients, and Chang et al. [23] investigate a Transformer-based approach to exploit breathing and speech information simultaneously.

This work focuses on the automatic detection of patients with COVID-19 in the context of the Second *Diagnosing COVID-19 using Acoustics* (DiCOVA) Challenge [24, 25]. We use the spectrogram representations of cough, breath, and speech samples to train neural networks composed of two main blocks: the first block aims at extracting embedded representations from the spectrograms, the second block is responsible for the actual classification. The embedded representations from the different sound types are extracted with dedicated *Convolutional Neural Networks* (CNNs). We explore the use of an outer product-based approach [26, 27] –to the best of the authors' knowledge for the first time in this task– to fuse the extracted representations with the goal to enrich the information for the final classification. Additionally, we also aim to investigate whether using the patients' sex as *a priori* information and introducing contextual attention mechanisms to the network can be beneficial for the task at hand.

The rest of the paper is laid out as follows. Section 2 describes the dataset analysed, while Section 3 details the methodology followed. Section 4 compiles and analyses the results obtained, and Section 5 concludes the paper.

## 2. Dataset

In this work, we use the dataset released as part of the Second DiCOVA Challenge [24, 25]. This dataset contains acoustic samples of COVID-19 positive and negative (healthy) patients from three different sound types produced by the human respiratory system; specifically, from coughs, breaths, and speech. Although the sampling rate of the acoustic samples provided is 44.1 kHz, an initial exploration of the dataset revealed the existence of samples without frequency content 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

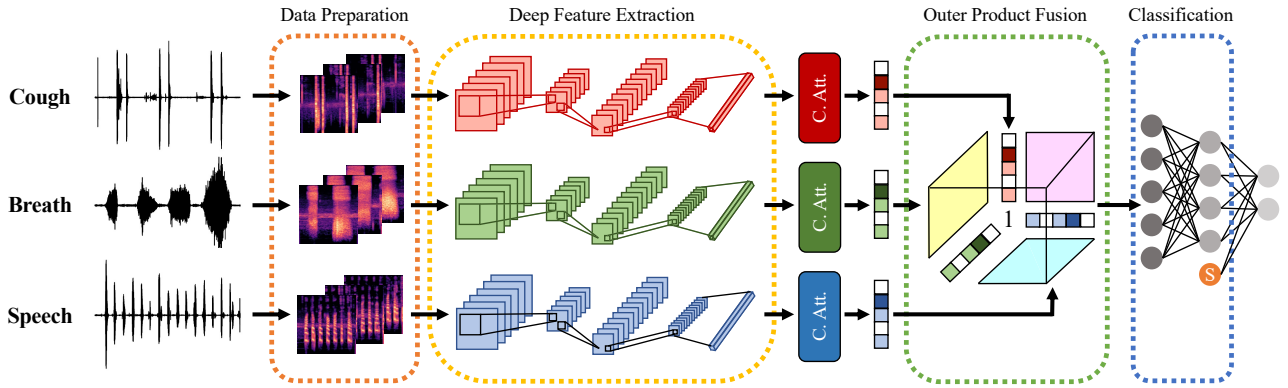


Figure 1: Diagram illustrating the system implemented, which receives a cough, breath, and speech sample as input, and outputs the probability to belong to a COVID-19 or a healthy patient. Patients’ sex is fed into the second layer of the classification block.

Table 1: 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).

|          | Validation | Test    | $\Sigma$ |
|----------|------------|---------|----------|
| 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  |
| $\Sigma$ | 10:15:00   | 4:30:23 | 14:45:23 |

Table 2: 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 | $\Sigma$ |
|----------|------------|----------|----------|------|----------|
|          | Positive   | Negative | $\Sigma$ |      |          |
| Females  | 53         | 189      | 242      | 119  | 361      |
| Males    | 119        | 604      | 723      | 352  | 1 075    |
| $\Sigma$ | 172        | 793      | 965      | 471  | 1 436    |

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 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 1). The dataset contains information from a total of 1 436 patients (cf. Table 2): 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.

### 3. Methodology

This section presents the methodology (cf. Figure 1). Section 3.1 details the pre-processing applied to the audio samples, Section 3.2 describes the networks implemented, and Section 3.3 summarises their training details.

#### 3.1. Data Preparation

The respiratory sounds are first downsampled to 16 kHz to overcome the disparity between recording devices, avoiding our networks to perform the COVID-19 detection based on the presence or the absence of frequency content in the upper frequencies of the spectrogram (cf. Section 2). The duration of the respiratory sounds recorded by each patient is heterogeneous. To assist in the fusion, we compute the longest sound from each patient and use this information to extend the shorter ones via repetition to homogenise their duration. Next, we window each respiratory sound separately into frames of 5 sec length with a 50 % overlap. We compute the magnitude of the *Short-Time Fourier Transform* (STFT) of each individual frame using a window length of 4 096 samples (256 ms), and a hop size of 128 samples (8 ms) to obtain its spectrogram representation. The spectrograms use a logarithmic frequency scale, and the *magma* colour map. Once normalised, each spectrogram is stored as a colour image of  $224 \times 224$  pixels.

The generated spectrograms are standardised before training. The standardisation parameters ( $\mu$  and  $\sigma$ ) are computed from all the spectrograms corresponding to the current sound type that belong to the training partition. To downsize the effect of the COVID-19 imbalanced data (cf. Table 2), we augment the spectrograms corresponding to the COVID-19 positive patients via replication. Despite considering other data augmentation strategies, such as filtering or additive noise, we decided not to alter the original samples in any way, as the relevant acoustic information for the task at hand is not clear yet. Replication may introduce redundancy in the training material, but we believe it can still be useful, as the number of positive and negative samples is significantly different.

#### 3.2. Models Description

This passage describes the network architectures implemented and investigated in this work.

##### 3.2.1. Baseline Models

The networks implemented are composed of two main blocks: the first block extracts deep learnt representations from the spectrograms of the cough ( $f_C$ ), breath ( $f_B$ ), and speech ( $f_S$ ) samples, while the second block performs the actual classification. For the feature extraction block, we compare two different architectures. The first architecture implements two convolutional layers with 16 and 4 filters, respectively, with a kernel size of  $3 \times 3$  and a stride of 1. Following each convolutional layer, we

use batch normalisation, and the output is transformed using a *Rectified Linear Unit* (ReLU) function. A 2-dimensional max pooling layer and a 2-dimensional adaptive average pooling layer are implemented at the end of the first and second convolutional block, respectively. This way, we force the output of the feature extraction block to produce 4 features per filter. The second architecture uses the ResNet18 model [28] without the last layer. Specifically, we use the pre-trained weights to initialise the network and fine-tune them during training for the task at hand. An additional linear layer is included in this architecture to reduce the dimensionality of the features from 512 to 16. When using the ResNet18 architecture, the spectrograms, represented as an RGB image, are standardised using the pre-defined parameters  $\mu = [0.485, 0.456, 0.406]$  and  $\sigma = [0.229, 0.224, 0.225]$ . The learnt features from both architectures have the same dimensionality, defined as a compromise with the dimensionality of the fused representation, and are finally flattened into a 1-dimensional representation.

The deep learnt representations from each sound type are extracted using a dedicated feature extraction block. In this work, we investigate the inner fusion of these embedded representations using an outer product-based approach [26, 27], which can be mathematically defined as:

$$\mathbf{f}_{C \otimes B \otimes S} = \begin{bmatrix} \mathbf{f}_C \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{f}_B \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{f}_S \\ 1 \end{bmatrix}. \quad (1)$$

When the three sound types are fused together, the outer product generates a cube with the following properties: i) the original representations are preserved in the edges of the cube, ii) each face of the cube contains information from the fusion of two sound types, and iii) the inner part of the cube fuses information from the three sound types all together. The fused representation is flattened before being fed into the final, classification block of the network. This fusion layer is implemented when training multi-type models, which combine at least two sound types, and omitted when training mono-type models, which consider a single sound type for predicting COVID-19.

The classification block of the network contains two fully connected layers, preceded by a dropout layer with probability 0.3. The number of input neurons in this block depends on the number of sound types to be fused. Nevertheless, the number of output neurons is fixed to 8. The output of this first layer is transformed using a ReLU activation function. The transformed representation is finally fed into the second layer of this block, which contains two output neurons with a Softmax activation function. This way, the outputs of the network can be interpreted as probability scores.

### 3.2.2. Sex-based Models

This model expands the baseline model described in Section 3.2.1 to consider the sex of the patients when inferring their COVID-19 status. Specifically, a binary encoded representation of the patient's sex is fed into the second layer of the classification block of the network. The number of input features to the classification block depends on the number of sound types to be fused. Introducing the sex information in the first layer of this block would complicate understanding if the performance of the network is conditioned by the patient's sex or by the number of input features. Thus, we opted for feeding this information into the second layer of the classification block, where the number of neurons corresponding to the sound representations is fixed.

### 3.2.3. Contextual Attention-based Models

This model also expands the baseline model described in Section 3.2.1, but, in this case, using a dedicated contextual attention mechanism at the output of each feature extraction block. The aim of this mechanism is to help highlight the salient information from the embedded representations learnt. Representing the embedded representations learnt as  $\mathbf{f}_N$ , where  $N \in \{C, B, S\}$  depending on the input sound type, the contextual attention mechanism is mathematically defined as:

$$\mathbf{u} = \tanh(\mathbf{W}\mathbf{f}_N + \mathbf{b}), \quad (2)$$

$$\boldsymbol{\alpha} = \frac{\exp(\mathbf{u}^T \mathbf{u}_c)}{\sum \exp(\mathbf{u}^T \mathbf{u}_c)}, \quad (3)$$

$$\tilde{\mathbf{f}}_N = \boldsymbol{\alpha} \mathbf{f}_N, \quad (4)$$

where  $\mathbf{W}$ ,  $\mathbf{b}$ , and  $\mathbf{u}_c$  are parameters to be learnt by the network. The parameter  $\mathbf{u}_c$  can be interpreted as the context vector. The attention-based representation obtained,  $\tilde{\mathbf{f}}_N$ , is then fed into the classification block of the network when training mono-type models, or fused when training multi-type models.

### 3.3. Networks Training

For a fair comparison among the models, these are all trained under the exact same conditions. We use 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  $\mathcal{L}_{AUC} = 1 - \text{AUC}$  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 defined by the Challenge organisers. Each fold is trained during a specific number of epochs. Hence, 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.

## 4. Experimental Results

The results obtained using specific CNNs and using ResNet18-based CNNs are summarised in Tables 3 and 4, respectively. One of the main insights from our experiments is that the fusion of breath and speech samples outperforms the multi-type models resulting from the combination of all other sound types and the mono-type models in 3 out of the 4, and in 2 out of the 4 scenarios investigated with the specific CNNs, and the ResNet18-based CNNs, respectively. Likewise, when we look at the mono-type models ( $C$ ,  $B$ ,  $S$ ), we observe that the models using the breath and the speech samples score higher results in comparison to the models using coughs only.

We observe that the mono-type models considering the patients' sex only improve the performance of the cough-based models, while they barely have an effect on the breath-based models. Patients' sex negatively impacts the performance of the speech-based models. Although there is no clear pattern to determine the suitability of considering patients' sex and/or using contextual attention, we note that the models surpassing the baseline with the specific CNNs use one of the three variants in most of the cases. The contextual attention-based model fusing

Table 3: AUC measurements (%) obtained from the mono- and multi-type models trained using specific CNNs for feature extraction (Baseline). These models consider the patient’s sex for the analysis (Sex), use contextual attention mechanisms (C. Att.), and their combination (Sex & C. Att.).

| Sound types             | Set  | Baseline     | Sex          | C. Att.      | Sex & C. Att. |
|-------------------------|------|--------------|--------------|--------------|---------------|
| $C$                     | Val. | 63.56        | 65.16        | 63.86        | 67.62         |
|                         | Test | 61.56        | 65.16        | 64.01        | 67.71         |
| $B$                     | Val. | 72.83        | 73.83        | 72.73        | 71.96         |
|                         | Test | 79.38        | 79.85        | 76.51        | 76.79         |
| $S$                     | Val. | 71.90        | 72.92        | 72.49        | 73.52         |
|                         | Test | 80.04        | 75.53        | 78.35        | 78.32         |
| $C \otimes B$           | Val. | 74.68        | 74.94        | 74.14        | 75.41         |
|                         | Test | 80.02        | 80.37        | –            | –             |
| $C \otimes S$           | Val. | 72.45        | 71.59        | 74.58        | 74.40         |
|                         | Test | –            | –            | –            | –             |
| $B \otimes S$           | Val. | 76.74        | 77.98        | 76.04        | 77.92         |
|                         | Test | <b>81.95</b> | <b>83.89</b> | <b>84.06</b> | 82.35         |
| $C \otimes B \otimes S$ | Val. | 72.91        | 73.71        | 78.22        | 76.77         |
|                         | Test | –            | –            | 81.98        | <b>83.25</b>  |

breath and speech samples obtains the best performance with an AUC of 84.06 %. With the ResNet18-based CNNs, the baseline models obtain the best AUC scores in most of the cases. The baseline model fusing breath and speech samples scores the best AUC of 84.26 %.

Although the transfer learning approach obtains the best performance, the specific CNNs obtain similar results with a simpler structure. Further experiments are needed to better understand the impact of patients’ sex in the fused scenarios, as we hypothesise it is downsized as a result of a magnitude difference between the sex representation and the deep learnt features at the intermediate layer of the classification block.

We finally compare in Table 5 the performance of our best mono- and multi-type models, obtained when using ResNet18-based CNNs for feature extraction (cf. Table 4), with the performance of the baseline models reported by the Challenge organisers [25]. The best AUC when using a mono-type model exploiting the cough samples is obtained with the sex-based model. When using only the breath and speech samples, our mono-type baseline models –which neither consider the patients’ sex, nor implement the contextual attention mechanism– obtain the highest AUC scores. In the multi-type scenario, our baseline model achieves the best AUC score when fusing the embedded representations learnt from the breath and speech samples.

As it can be observed from the results reported in Table 5, neither our mono-type, nor our multi-type models surpass the Challenge baseline models. Nevertheless, the difference in terms of the AUC on the test set between the multi-type models is lower than 1 %, which suggests a similar performance. While we extract log-spectrogram representations from the input signals, the Challenge baseline uses 64 mel-spectrogram features, expanded with their first and second order temporal derivatives. It is therefore reasonable to argue about the suitability of the log-spectrogram representations for this task. The multi-type model proposed by the Challenge organisers averages the probabilities obtained with their mono-type models, and, as a consequence, does not exploit the complementarity among the involved sound types for training their models. This result casts doubt on the efficacy of the proposed outer product-based fusion. However, the comparison is not fair, as the inferior performance could be caused by the different nature of the input features used. Thus,

Table 4: AUC measurements (%) obtained from the mono- and multi-type models trained using ResNet18-based CNNs for feature extraction (Baseline). These models consider the patient’s sex for the analysis (Sex), use contextual attention mechanisms (C. Att.), and their combination (Sex & C. Att.).

| Sound types             | Set  | Baseline     | Sex          | C. Att.      | Sex & C. Att. |
|-------------------------|------|--------------|--------------|--------------|---------------|
| $C$                     | Val. | 76.42        | 74.48        | 73.12        | 73.39         |
|                         | Test | 64.69        | 68.76        | 66.60        | 68.15         |
| $B$                     | Val. | 78.78        | 79.16        | 78.62        | 80.78         |
|                         | Test | 80.35        | <b>79.91</b> | 77.77        | 80.21         |
| $S$                     | Val. | 79.02        | 79.25        | 78.19        | 79.10         |
|                         | Test | 81.86        | 75.21        | 78.89        | <b>81.66</b>  |
| $C \otimes B$           | Val. | 76.35        | 76.79        | 74.46        | 72.48         |
|                         | Test | –            | 75.03        | –            | –             |
| $C \otimes S$           | Val. | 75.56        | 76.19        | 77.69        | 76.64         |
|                         | Test | –            | –            | 77.07        | –             |
| $B \otimes S$           | Val. | 78.06        | 79.87        | 80.12        | 77.65         |
|                         | Test | <b>84.26</b> | 73.48        | <b>83.63</b> | 81.48         |
| $C \otimes B \otimes S$ | Val. | 76.90        | 75.84        | 76.79        | 76.07         |
|                         | Test | 76.78        | –            | –            | –             |

Table 5: AUC measurements (%) defined as the Challenge baselines per sound type compared with the models presented in this work that obtained the highest performances on the test set.

|                         | Set  | Mono-Type    |              |              | Multi-Type   |
|-------------------------|------|--------------|--------------|--------------|--------------|
|                         |      | $C$          | $B$          | $S$          |              |
| Challenge Baseline [25] | Val. | 75.21        | 77.25        | 80.16        | 81.67        |
|                         | Test | <b>74.89</b> | <b>84.50</b> | <b>84.26</b> | <b>84.70</b> |
| Our Optimal Models      | Val. | 74.48        | 78.78        | 79.02        | 78.06        |
|                         | Test | 68.76        | 80.35        | 81.86        | 84.26        |

the assessment of the proposed outer product-based approach using the input features extracted in the Challenge baseline is necessary to fairly compare the performance of the multi-type models and draw valid conclusions.

## 5. Conclusions

We proposed an outer product-based approach to fuse the embedded representations extracted from cough, breath, and speech samples towards the detection of patients with COVID-19. The results obtained indicated the suitability of fusing the embedded representations learnt from the spectrograms of breath and speech samples for this task. The best models fusing these sound types scored an AUC of 84.06 % and 84.26 % with specific and ResNet18-based CNNs for extracting deep learnt features from the spectrograms, respectively.

In addition to exploring other fusion techniques, we would like to carry on this research by applying dedicated processing pipelines to each sound type. Furthermore, we hypothesise whether transforming the spectrograms with contextual attention mechanisms might contribute to better understand the acoustic information relevant to detect COVID-19.

## 6. Acknowledgements

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 Spanish Ministry of Science and Innovation under the Musical AI project (PID2019-111403GB-I00).

## 7. References

- [1] S. Amiriparian, S. Pugachevskiy, N. Cummins, S. Hantke, J. Pohjalainen, G. Keren, and B. Schuller, "CAST a database: Rapid targeted large-scale big data acquisition via small-world modelling of social media platforms," in *Proc. of the 7th Intl. Conference on Affective Computing and Intelligent Interaction*. San Antonio, TX, USA: IEEE, 2017, pp. 340–345.
- [2] A. Mallol-Ragolta, Z. Zhao, L. Stappen, N. Cummins, and B. W. Schuller, "A Hierarchical Attention Network-Based Approach for Depression Detection from Transcribed Clinical Interviews," in *Proc. of Interspeech*. Graz, Austria: ISCA, 2019, pp. 221–225.
- [3] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, and M. 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: ACM, 2019, pp. 3–12.
- [4] A. Mallol-Ragolta, S. Dhamija, and T. E. Boulton, "A Multimodal Approach for Predicting Changes in PTSD Symptom Severity," in *Proc. of the 20th Intl. Conference on Multimodal Interaction*. Boulder, CO, USA: ACM, 2018, pp. 324–333.
- [5] S. Heidarian, P. Afshar, A. Mohammadi, M. J. Rafiee, A. Oikonomou, K. N. Plataniotis, and F. Naderkhani, "CT-CAPS: Feature Extraction-Based Automated Framework for COVID-19 Disease Identification From Chest CT Scans Using Capsule Networks," in *Proc. of the ICASSP*. Toronto, ON, Canada: IEEE, 2021, pp. 1040–1044.
- [6] P. Garg, R. Ranjan, K. Upadhyay, M. Agrawal, and D. Deepak, "Multi-Scale Residual Network for COVID-19 Diagnosis Using CT-Scans," in *Proc. of the ICASSP*. Toronto, ON, Canada: IEEE, 2021, pp. 8558–8562.
- [7] S. Rajaraman, J. Siegelman, P. O. Alderson, L. S. Folio, L. R. Folio, and S. K. Antani, "Iteratively Pruned Deep Learning Ensembles for COVID-19 Detection in Chest X-Rays," *IEEE Access*, vol. 8, pp. 115 041–115 050, 2020.
- [8] A. Mallol-Ragolta, H. Cuesta, E. Gómez, and B. W. Schuller, "Cough-Based COVID-19 Detection with Contextual Attention Convolutional Neural Networks and Gender Information," in *Proc. of Interspeech*. Brno, Czechia: ISCA, 2021, pp. 941–945.
- [9] V. Karas and B. W. Schuller, "Recognising COVID-19 from Coughing Using Ensembles of SVMs and LSTMs with Handcrafted and Deep Audio Features," in *Proc. of Interspeech*. Brno, Czechia: ISCA, 2021, pp. 911–915.
- [10] R. Solera-Ureña, C. Botelho, F. Teixeira, T. Rolland, A. Abad, and I. Trancoso, "Transfer Learning-Based Cough Representations for Automatic Detection of COVID-19," in *Proc. of Interspeech*. Brno, Czechia: ISCA, 2021, pp. 436–440.
- [11] 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 Intl. Conference on Knowledge Discovery & Data Mining*. Virtual Conference: ACM, 2020, pp. 3474–3484.
- [12] H. Coppock, A. Gaskell, P. Tzirakis, A. Baird, L. Jones, and B. Schuller, "End-to-end convolutional neural network enables COVID-19 detection from breath and cough audio: a pilot study," *BMJ Innovations*, vol. 7, no. 2, pp. 356–362, 2021.
- [13] S. Deshmukh, M. Al Ismail, and R. Singh, "Interpreting Glottal Flow Dynamics for Detecting COVID-19 From Voice," in *Proc. of the ICASSP*. Toronto, ON, Canada: IEEE, 2021, pp. 1055–1059.
- [14] J. Han, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring Automatic COVID-19 Diagnosis via Voice and Symptoms from Crowdsourced Data," in *Proc. of the ICASSP*. Toronto, ON, Canada: IEEE, 2021, pp. 8328–8332.
- [15] R. Solera-Ureña, C. Botelho, F. Teixeira, T. Rolland, A. Abad, and I. Trancoso, "Transfer Learning-Based Cough Representations for Automatic Detection of COVID-19," in *Proc. of Interspeech*. Brno, Czechia: ISCA, 2021, pp. 436–440.
- [16] I. Södergren, M. P. Nodeh, P. C. Chhipa, K. Nikolaidou, and G. Kovács, "Detecting COVID-19 from Audio Recording of Coughs Using Random Forests and Support Vector Machines," in *Proc. of Interspeech*. Brno, Czechia: ISCA, 2021, pp. 916–920.
- [17] J. Harvill, Y. R. Wani, M. Hasegawa-Johnson, N. Ahuja, D. Beiser, and D. Chestek, "Classification of COVID-19 from Cough Using Autoregressive Predictive Coding Pretraining and Spectral Data Augmentation," in *Proc. of Interspeech*. Brno, Czechia: ISCA, 2021, pp. 926–930.
- [18] S. Bhosale, U. Tiwari, R. Chakraborty, and S. K. Koppurapu, "Contrastive Learning of Cough Descriptors for Automatic COVID-19 Preliminary Diagnosis," in *Proc. of Interspeech*. Brno, Czechia: ISCA, 2021, pp. 946–950.
- [19] B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen, S. Ottl, M. Gerczuk, P. Tzirakis, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, L. J. Rothkrantz, J. A. Zwarts, J. Treep, and C. S. Kaandorp, "The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primitives," in *Proc. of Interspeech*. Brno, Czechia: ISCA, 2021, pp. 431–435.
- [20] E. Casanova, A. Candido Jr., R. C. Fernandes Jr., M. Finger, L. R. S. Gris, M. A. Ponti, and D. P. Pinto da Silva, "Transfer Learning and Data Augmentation Techniques to the COVID-19 Identification Tasks in ComParE 2021," in *Proc. of Interspeech*. Brno, Czechia: ISCA, 2021, pp. 446–450.
- [21] M. A. Nessiem, M. M. Mohamed, H. Coppock, A. Gaskell, and B. W. Schuller, "Detecting COVID-19 from Breathing and Coughing Sounds using Deep Neural Networks," in *Proc. of the 34th Intl. Symposium on Computer-Based Medical Systems*. Aveiro, Portugal: IEEE, 2021, pp. 183–188.
- [22] S. Liu, A. Mallol-Ragolta, and B. W. Schuller, "COVID-19 Detection with a Novel Multi-Type Deep Fusion Method using Breathing and Coughing Information," in *Proc. of the 43rd Intl. Conference of the Engineering in Medicine Biology Society*. Guadalajara, Mexico – Virtual Event: IEEE, 2021, pp. 1840–1843.
- [23] Y. Chang, Z. Ren, and B. W. Schuller, "Transformer-based CNNs: Mining Temporal Context Information for Multi-sound COVID-19 Diagnosis," in *Proc. of the 43rd Intl. Conference of the Engineering in Medicine Biology Society*. Guadalajara, Mexico – Virtual Event: IEEE, 2021, pp. 2335–2338.
- [24] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. R. Chetupalli, N. R., P. K. Ghosh, and S. Ganapathy, "Coswara – A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis," in *Proc. of Interspeech*. Shanghai, China: ISCA, 2020, pp. 4811–4815.
- [25] N. K. Sharma, S. R. Chetupalli, D. Bhattacharya, D. Dutta, P. Mote, and S. Ganapathy, "The Second DiCOVA Challenge: Dataset and Performance Analysis for COVID-19 Diagnosis using Acoustics," in *Proc. of the Intl. Conference on Acoustics Speech Signal Processing*. Singapore: IEEE, 2022, to appear.
- [26] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor Fusion Network for Multimodal Sentiment Analysis," in *Proc. of the Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: ACL, 2017, pp. 1103–1114.
- [27] A. Mallol-Ragolta, A. Semertzidou, M. Pateraki, and B. Schuller, "Outer Product-Based Fusion of Smartwatch Sensor Data for Human Activity Recognition," *Frontiers in Computer Science, section Mobile and Ubiquitous Computing*, vol. 4, pp. 1–10, March 2022, article ID 796866.
- [28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. of the Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA: IEEE, 2016, pp. 770–778.