INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Cough-based COVID-19 Detection with Contextual Attention Convolutional Neural Networks and Gender Information

Adria Mallol-Ragolta[1]*, Helena Cuesta*[2]*, Emilia Gómez*[2,3]*, and Björn W. Schuller*[1,4]

[1] EIHW – Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany
[2] MTG – Music Technology Group, Universitat Pompeu Fabra, Spain
[3] Joint Research Centre, European Commission, Spain
[4] GLAM – Group on Language, Audio & Music, Imperial College London, UK

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

Abstract

The aim of this contribution is to automatically detect COVID-19 patients by analysing the acoustic information embedded in coughs. COVID-19 affects the respiratory system, and, consequently, respiratory-related signals have the potential to contain salient information for the task at hand. We focus on analysing the spectrogram representations of cough samples with the aim to investigate whether COVID-19 alters the frequency content of these signals. Furthermore, this work also assesses the impact of gender in the automatic detection of COVID-19. To extract deep-learnt representations of the spectrograms, we compare the performance of a cough-specific, and a Resnet18 pre-trained Convolutional Neural Network (CNN). Additionally, our approach explores the use of contextual attention, so the model can learn to highlight the most relevant deep-learnt features extracted by the CNN. We conduct our experiments on the dataset released for the Cough Sound Track of the DiCOVA 2021 Challenge. The best performance on the test set is obtained using the Resnet18 pre-trained CNN with contextual attention, which scored an Area Under the Curve (AUC) of 70.91 % at 80 % sensitivity.

Index Terms: COVID-19, acoustics, machine learning, respiratory diagnosis, healthcare

1. Introduction

The outbreak of the *Coronavirus Disease 2019* (COVID-19) has dramatically stressed the health systems worldwide. At the time of writing, the *World Health Organization* (WHO) reports more than 175.3 M confirmed cases, and more than 3.7 M confirmed deaths of COVID-19 across the globe. Despite the vaccines, massive population screenings will still be needed to control the spread of this disease, and its strains. Current medical diagnostic tools are time consuming, and burden public expenditures. Thus, there is an opportunity to develop new digital, diagnostic tools to improve the monitoring, and the early detection of COVID-19 at a large scale cost-effectively.

One of the core elements for effective digital health solutions is the use of *Artificial Intelligence* (AI). AI-based systems have been successfully employed to detect coughs or sneezes [1], or to analyse breath signals [2], among others. Furthermore, AI has also been used in the field of mental health, providing solutions to recognise mental illnesses, such as depression [3, 4, 5] or *Post-Traumatic Stress Disorder* (PTSD) [6]. The current context of the pandemic challenges researchers to focus on the development of automatic COVID-19 detection tools.

The symptomatology of COVID-19 presents affectations in the respiratory system. In this direction, the research community has already started investigating the use of AI techniques to analyse lung-based information through chest X-ray images [7, 8, 9, 10, 11] or CT scans [12, 13, 14]. Moreover, related works in the literature explored the use of respiratory-related body signals under the assumption that the acoustics of these signals have a high potential to contain salient information to diagnose COVID-19 patients. The signals considered include breaths [15, 16], coughs [17, 18], and even speech [19, 20].

This paper presents our contribution to the Cough Sound Track of the *Diagnosing COVID-19 using Acoustics* (DiCOVA) 2021 Challenge [21]. We opt for analysing the spectrogram representations of the cough signals with the aim to i) investigate whether COVID-19 symptomatology alters the frequency content of coughs, and ii) assess the impact of gender[1] in the automatic detection of COVID-19 patients. Our approach relies on *Convolutional Neural Networks* (CNNs) to extract salient information from the spectrograms, combined with *Fully Connected* (FC) layers responsible for the classification of the embedded features learned. Our approach also explores the use of contextual attention, so the network learns to highlight the most relevant embedded features 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 from the experiments performed, and Section 5 concludes the paper and suggests potential future work directions.

2. Dataset

The dataset used in this work was released for the Cough Sound Track of the DiCOVA Challenge 2021 [21]. This dataset consists of cough sounds recorded from COVID-19 positive and non-COVID-19 (healthy) patients, as well as their associated metadata, i. e., COVID-19 status (positive or negative), gender, and nationality. No information about symptomatic/asymptomatic COVID-19 positive patients is provided in the dataset. The cough recordings are split into a training and a test partition. The former contains the ground truth COVID-19 status, while the latter is blind to the participants. To assess the performance of the models, the Challenge organisers require following a 5-fold cross-validation method, and distribute the recordings that belong to each fold.

The training partition is composed of 1 040 audio recordings of different durations, ranging from ca. 1 sec up to 15 sec, with an average duration of 4.7 sec. Male and female patients recorded 791 and 249 samples, respectively. The test partition contains a total of 233 audio recordings: 171 and 62 from male and female

[1] We use the term *gender* for consistency with the nomenclature used in the baseline paper.

http://dx.doi.org/10.21437/Interspeech.2021-1052

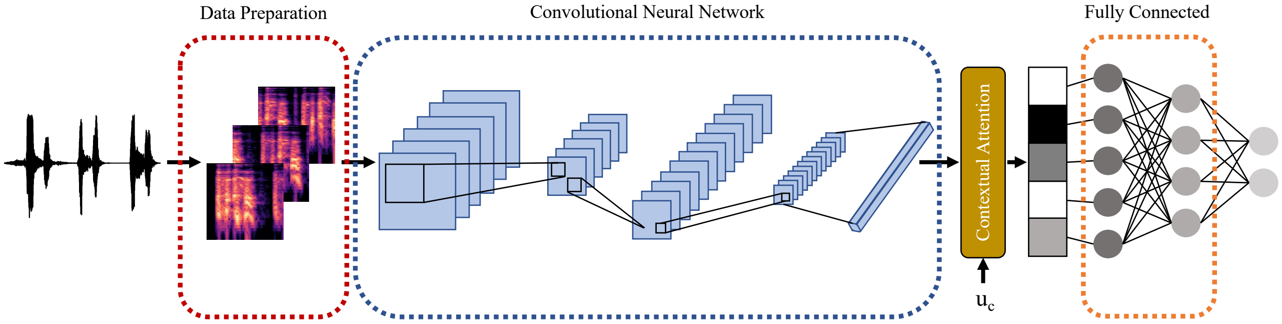

Figure 1: *Diagram illustrating the system implemented, which receives a cough sample as input, and outputs the probability of the input sample to correspond to a COVID-19 or a healthy patient. The feature extraction of the segmented spectrograms is performed with a convolutional neural network. The most relevant embedded features are highlighted using a contextual attention mechanism, before the final classification using two stacked fully connected layers.*

patients, respectively. The duration of the audio recordings ranges from 1 sec to 12 sec, with an average of 4.6 sec. The training dataset is imbalanced in terms of positive and negative examples: from the 1 040 samples, 965 are negative (healthy patients), while only 75 are positive (COVID-19 patients).

The provided audio files are sampled at 44.1 kHz. Our preliminary spectral analysis of a subset of the recordings revealed that a substantial amount of them does not have any frequency content above 8 kHz. We hypothesise a potential reason to explain this is the use of low-quality equipment by patients when recording their cough samples, e. g. , with mobile devices. Therefore, we resample all audio files to a common sampling rate of 8 kHz to account for the diversity of devices used for recording. Besides, the lack of frequency content above 8 kHz results in a dark patch in the spectrogram representations of the corresponding samples, which creates noise in the training data.

3. Methodology

This section describes the methodology, illustrated in Figure 1. Section 3.1 details the preprocessing applied to the cough samples, Section 3.2 introduces the models implemented, and Section 3.3 summarises the parameters used to train the networks.

3.1. Data Preparation

This section details the data preparation stage of our approach, which has several steps: silence removal, feature extraction, data patch generation, and data augmentation.

3.1.1. Sound Activity Detection

Each audio sample in the dataset contains a sequence of coughs. A short amount of silence separates consecutive cough samples within each sequence. We consider these silent regions to be irrelevant in the detection of COVID-19, and, therefore, we use a *Sound Activity Detector* (SAD) to filter them out. After the resampling step, the audio files are subsequently passed through a SAD based on the *Root-Mean-Square* (RMS) value of the audio samples in the time domain [22, 23, 24]. We compute the RMS using the `librosa` Python library [25] and a frame length of 64 msec. We use min-max normalisation to scale each audio file's RMS, and we discard all frames below a threshold of 0.1 (set empirically). After the SAD step, we concatenate all frames above the threshold, and save the result as a new audio file for further processing. As an additional experiment, we compared the RMS-based SAD to a SAD based on spectral flux [26, 27], which detects abrupt changes in the spectral domain. Although cough is an example of such a change, the preliminary exploration of the results using both methods showed that the RMS-based SAD worked better in this context. Note that to assess the effectiveness of silence removal in the detection of COVID-19 patients, our experiments use both the original and the cough-only audio files. Details about the models trained, and the results obtained are given in Section 3.2 and Section 4, respectively.

3.1.2. Feature Extraction and Patch Generation

Our approach uses the spectrogram as the input representation of the cough samples. We use the *Short-Time Fourier Transform* (STFT) function from `librosa` to calculate the spectrogram of each cough sample in the dataset. We use a window size of 1 024 samples (128 msec), and a hop length of 128 samples (16 msec). With this configuration, we extract the spectrograms using different parameters to compare their impact on the final results: we compare the spectrograms with a linear or logarithmic frequency scale, and two different colour maps, namely, *viridis* and *magma*. The colour map parameter is especially relevant, because spectrograms are exported as images of 256 × 256 pixels for further use. The preliminary experiments conducted to assess the impact of these parameters, not reported in this work, were not conclusive enough. Nonetheless, analysing the trends in the results obtained, we decided to focus our investigation on the spectrogram representations of the cough samples using the logarithmic frequency scale, and *magma* as the colour map. To overcome the different duration of the samples in the dataset, we fix the length of the cough samples to be fed into the models. We decide modelling the cough samples using acoustic frames of 1 sec length. Hence, the last step of the data preparation stage is the segmentation of all spectrograms into 1 sec length patches with a 50 % overlap. With this strategy, several patches from a single cough sample are used for training the models.

3.1.3. Data Augmentation

To address the imbalance between positive and negative examples (cf. Section 2), which impacts the number of spectrograms generated from the patches of the cough samples defined, we use data augmentation. Specifically, we increase the number of positive examples via replication, i. e. , including copies of the positive spectrograms to balance the dataset. We considered other forms of augmentation, such as filtering or additive noise. However, since it is not clear yet which kind of information from the audio is relevant for the task at hand, we decided not to alter the acoustic content in any way. Although replication introduces redundancy in the training set, we believe it is useful when the number of positive and negative examples differs significantly.

Table 1: *Summary comparing the performance of the models trained with a cough-specific CNN. The results reported on the validation partition are computed by averaging the performances obtained in each individual fold. The model that scored the highest AUC is highlighted.*

Models	Audio Files	Partition	AUC [%]	Sensitivity [%]	Specificity [%]
Baseline	Original	Val.	63.14	81.60	34.09
		Test	54.31	80.49	–
	Cough-only	Val.	62.86	81.60	38.65
		Test	58.31	80.49	–
Gender Based	Original	Val.	64.88	80.80	39.90
		Test	52.88	80.49	–
	Cough-only	Val.	65.32	82.40	38.86
		Test	**59.04**	80.49	–
Gender Specific	Original	Val.	58.91	83.20	33.89
		Test	52.88	80.49	–
	Cough-only	Val.	62.86	80.80	42.59
		Test	49.92	80.49	–

3.2. Models Description

This section presents the neural networks used to model the cough samples to detect COVID-19 patients. While Section 3.2.1 describes the architecture of the networks implemented, Section 3.2.2 details the procedure for our networks to consider gender information.

3.2.1. Network Architectures

The networks trained to detect COVID-19 from cough samples are composed of a first block, extracting embedded representations from the input spectrograms, and a second block, focusing on the classification of the embedded features depending on whether they belong to healthy or COVID-19 patients. For the latter, we employ two FC layers with 128 and 2 output neurons, respectively. While the first layer uses a *Rectified Linear Unit* (ReLU) as the activation function, the second one uses Softmax, so the network outputs can be interpreted as probability scores.

As our networks' inputs are spectrograms, the extraction of the embedded representations is implemented using CNNs. Specifically, we compare the performance of a cough-specific CNN trained from scratch with the performance of a ResNet18 pre-trained CNN [28]. The cough-specific CNN is implemented with two convolutional blocks with 32 and 64 channels, respectively, a square kernel of 3×3, and a stride of 1. Both blocks implement batch normalisation, and use ReLU as the activation function. While the first block includes a 2×2 max-pooling, the second one uses adaptive average pooling, so the learnt feature map has a dimension of 2×2.

To highlight the salient information from the embedded representations learnt, we include a contextual attention mechanism (adapted from [29] and [4]) between the two blocks of the network. Representing the deep features learnt as $\boldsymbol{h}$, the contextual attention mechanism is mathematically defined as follows:

$$\boldsymbol{u} = \tanh(\mathbf{W}\boldsymbol{h} + \mathbf{b}), \tag{1}$$

$$\boldsymbol{\alpha} = \frac{\exp\left(\boldsymbol{u}^T \mathbf{u_c}\right)}{\sum \exp\left(\boldsymbol{u}^T \mathbf{u_c}\right)}, \tag{2}$$

$$\tilde{\boldsymbol{h}} = \boldsymbol{\alpha}\boldsymbol{h}, \tag{3}$$

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 $\tilde{\boldsymbol{h}}$ is then fed into the FC layers for classification.

Table 2: *Summary comparing the performance of the models trained with a cough-specific CNN and contextual attention. The results reported on the validation partition are computed by averaging the performances obtained in each individual fold. The model that scored the highest AUC is highlighted.*

Models	Audio Files	Partition	AUC [%]	Sensitivity [%]	Specificity [%]
Baseline	Original	Val.	60.08	83.20	31.19
		Test	56.22	80.49	–
	Cough-only	Val.	63.00	83.20	34.61
		Test	58.34	80.49	–
Gender Based	Original	Val.	65.52	81.60	45.49
		Test	53.69	80.49	–
	Cough-only	Val.	63.39	84.00	39.17
		Test	**61.62**	80.49	–
Gender Specific	Original	Val.	58.08	81.60	34.40
		Test	50.77	80.49	–
	Cough-only	Val.	62.21	84.00	39.59
		Test	60.87	80.49	–

3.2.2. Gender Awareness

Assessing the impact of gender in the automatic detection of COVID-19 patients is also one of this work's goals. To address this, we explore three different network configurations. The first one does not consider any gender information, and is used as a baseline for our experiments. The second one, referred to as gender-based models in our experiments, includes an encoded representation of the patients' gender, which is concatenated with the deep-learnt features extracted. Both are fed into the FC layers of the network. The third and last configuration, referred to as gender-specific models in our experiments, trains gender-specific models, so female and male coughs are analysed with models trained using samples from patients of the same gender.

3.3. Networks Training

All models are trained to minimise the Categorical Cross-Entropy Loss, using Adam as the optimiser with a fixed learning rate of $1e^{-3}$. Network parameters are updated in batches of 32 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 ten consecutive epochs. To assess the models, we follow a 5-fold cross-validation approach, as defined by the challenge organisers. Each fold is trained during a specific number of epochs. Therefore, when modelling all training material and to prevent overfitting, the training epochs are determined by computing the median of the training epochs processed in each fold.

4. Experimental Results

Our models estimate the probability of the input cough to correspond to a COVID-19 patient. Using these probabilities and a set of thresholds between 0 and 1, we can compute the *Receiver Operating Characteristic* (ROC) curve. The ROC curve plots the evolution of the *True Positive Rate* (TPR) against the *False Positive Rate* (FPR). The TPR, also referred to as sensitivity, corresponds to the percentage of positive examples correctly identified, i. e., the *True Positives* (TP). The FPR refers to the percentage of negative examples identified as positive, i. e., the *False Positives* (FP). Using the ROC curve, we quantify the models' performance using the *Area Under the Curve* (AUC) as our primary evaluation metric. We fix the model sensitivity at 80 %, and compute the model specificity as an additional measure of

Table 3: *Summary comparing the performance of the models trained with the pre-trained Resnet18 CNN. The results reported on the validation partition are computed by averaging the performances obtained in each individual fold. The model that scored the highest AUC is highlighted.*

Models	Audio Files	Partition	AUC [%]	Sensitivity [%]	Specificity [%]
Baseline	Original	Val.	62.32	81.60	33.16
		Test	68.43	80.49	–
	Cough-only	Val.	53.55	81.60	16.48
		Test	54.55	80.49	–
Gender Based	Original	Val.	64.35	82.40	38.24
		Test	**68.95**	80.49	–
	Cough-only	Val.	55.33	84.00	18.45
		Test	52.31	82.93	–
Gender Specific	Original	Val.	60.00	84.00	27.05
		Test	51.94	80.49	–
	Cough-only	Val.	57.58	81.60	26.53
		Test	58.62	97.56	–

Table 4: *Summary comparing the performance of the models trained with the pre-trained Resnet18 CNN and contextual attention. The results reported on the validation partition are computed by averaging the performances obtained in each individual fold. The model that scored the highest AUC is highlighted.*

Models	Audio Files	Partition	AUC [%]	Sensitivity [%]	Specificity [%]
Baseline	Original	Val.	62.97	81.60	37.41
		Test	69.17	80.49	–
	Cough-only	Val.	54.93	82.40	21.55
		Test	52.56	80.49	–
Gender Based	Original	Val.	61.59	80.80	31.19
		Test	69.89	80.49	–
	Cough-only	Val.	55.38	80.80	24.56
		Test	54.34	80.49	–
Gender Specific	Original	Val.	62.01	83.20	36.27
		Test	**70.91**	80.49	–
	Cough-only	Val.	59.56	81.60	35.23
		Test	59.01	80.49	–

the models' performance. The specificity, also known as the *True Negative Rate* (TNR), indicates the proportion of negative examples which are correctly identified, i. e., the *True Negatives* (TN). The validation results are determined by averaging the individual metrics computed for each fold.

As described in Section 3.1.2, several fixed-length spectrograms can be extracted from a single cough sample. Thus, at inference time, several probability scores can be predicted for a single sample. To overcome this issue, we compute the probability of a specific sample to belong to a COVID-19 patient as the median of the probabilities inferred from the corresponding spectrograms. The results obtained when assessing the models trained using a cough-specific CNN without and with contextual attention are summarised in Tables 1 and 2, respectively. The results obtained when assessing the models trained using the pre-trained ResNet18 CNN without and with contextual attention are compiled in Tables 3 and 4, respectively.

One of our experiments' main insights is that models that incorporate gender information outperform the baseline model in most of the cases. In this task, the gender of the patient is especially relevant: the vocal apparatus of females and males has a different shape and size, which results in significant differences both in the timbre and frequency range of the respiratory-related signals. We obtained the best performance with the gender-based model in three of the four scenarios investigated. Besides, although in Table 4 the gender-specific model achieved the highest AUC on the test partition, the corresponding AUC for the gender-based model is only 1 % lower, suggesting an equivalent performance. Further gender-focused evaluations using the model with the best performance on the test partition reported an AUC of 67.98 % and 50.00 % on the validation partition for male and female patients, respectively, highlighting the relevance of gender in this task. Hence, intuitively, gender-specific models should work better, but these cannot be fairly studied because of the imbalance of the data in terms of gender.

When we compare the results between cough-only and original audio files, we observe a clear difference: interestingly, the best performances using cough-only audio files on the test set were obtained with the cough-specific CNN (cf. Tables 1 and 2), while the original audio files scored the highest AUC using the pre-trained Resnet18 CNN (cf. Tables 3 and 4). One potential reason behind these differences is that ResNet18 is a pre-trained network for image classification, and not directly related to acoustics. In general, images can be quite heterogeneous, i. e., they have several elements, separated by edges, with higher gradients to be captured by the CNN. In our spectrogram images, part of the edges appear at the beginning and end of the silent regions because of the abrupt change in the frequency content. The cough-only audio files do not contain these silent regions, and, therefore, they become much more homogeneous. In contrast, the original audio files have more edges, and they would be more suitable for such a pre-trained network.

5. Conclusions

This work presented our contribution to the Cough Sound Track of the DiCOVA Challenge 2021, which addressed the automatic detection of COVID-19 patients from cough samples. Emphasising on the impact of gender, our approach focused on the extraction of deep features from the spectrogram representations of coughs using CNNs in combination with a contextual attention mechanism. Specifically, we compared the performance of a cough-specific CNN, and a pre-trained ResNet18 CNN.

A gender-specific pre-trained Resnet18 CNN with contextual attention scored the highest performance on the test set, with an AUC of 70.91 %. Globally, the obtained results support the use of gender-based models, highlighting the impact of gender in the detection of COVID-19 from coughs. The best cough-specific CNNs exploiting cough-only audio files achieved an AUC of 59.04 % and 61.62 % without and with contextual attention, respectively. The best pre-trained Resnet18 CNNs exploiting the original audio files obtained an AUC of 68.95 % and 69.89 % without and with contextual attention, respectively.

As future work, other state-of-the-art pre-trained CNNs in the computer vision or computer audition domains could be investigated to extract deep features from the spectrograms. Further research could also explore deeper cough-specific CNNs to extract more relevant deep features. Regardless of the technology, the medical research in the symptomatology of COVID-19 will provide valuable insights to develop more effective systems.

6. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No. 826506 (sustAGE) and No. 770376 (TROMPA). Further funding has been received from the FI Predoctoral Grant 2018FI-B01015 from AGAUR, Generalitat de Catalunya.

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* 7^{th} *International Conference on Affective Computing and Intelligent Interaction*. San Antonio, TX, USA: IEEE, 2017, pp. 340–345.

[2] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks," in *Proc. of Interspeech*. Shanghai, China: ISCA, 2020, pp. 2042–2046.

[3] T. Al Hanai, M. Ghassemi, and J. Glass, "Detecting Depression with Audio/Text Sequence Modeling of Interviews," in *Proc. of Interspeech*. Hyderabad, India: ISCA, 2018, pp. 1716–1720.

[4] 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.

[5] 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* 9^{th} *International Audio/Visual Emotion Challenge and Workshop*. Nice, France: ACM, 2019, pp. 3–12.

[6] A. Mallol-Ragolta, S. Dhamija, and T. E. Boult, "A Multimodal Approach for Predicting Changes in PTSD Symptom Severity," in *Proc. of the* 20^{th} *International Conference on Multimodal Interaction*. Boulder, CO, USA: ACM, 2018, pp. 324–333.

[7] H. Panwar, P. Gupta, M. K. Siddiqui, R. Morales-Menendez, and V. Singh, "Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet," *Chaos, Solitons & Fractals*, vol. 138, p. 109944, 2020.

[8] A. M. Ismael and A. Şengür, "Deep learning approaches for COVID-19 detection based on chest X-ray images," *Expert Systems with Applications*, vol. 164, p. 114054, 2021.

[9] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105608, 2020.

[10] 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.

[11] G. Jain, D. Mittal, D. Thakur, and M. K. Mittal, "A deep learning approach to detect Covid-19 coronavirus with X-Ray images," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 4, pp. 1391–1405, 2020.

[12] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Explainable COVID-19 Detection Using Chest CT Scans and Deep Learning," *Sensors*, vol. 21, no. 2, p. 455, 2021.

[13] P. Gifani, A. Shalbaf, and M. Vafaeezadeh, "Automated detection of COVID-19 using ensemble of transfer learning with deep convolutional neural network based on CT scans," *International Journal of Computer Assisted Radiology and Surgery*, vol. 16, pp. 115–123, 2021.

[14] A. Mohammed, C. Wang, M. Zhao, M. Ullah, R. Naseem, H. Wang, M. Pedersen, and F. A. Cheikh, "Weakly-Supervised Network for Detection of COVID-19 in Chest CT Scans," *IEEE Access*, vol. 8, pp. 155 987–156 000, 2020.

[15] 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.

[16] 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 *Proc. of the* 26^{th} *International Conference on Knowledge Discovery & Data Mining*. Virtual Conference: ACM, 2020, pp. 3474–3484.

[17] A. Imran, I. Posokhova, H. N. Qureshi, U. Masood, M. S. Riaz, K. Ali, C. N. John, M. I. Hussain, and M. Nabeel, "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app," *Informatics in Medicine Unlocked*, vol. 20, p. 100378, 2020.

[18] M. Cohen-McFarlane, R. Goubran, and F. Knoefel, "Novel Coronavirus Cough Database: NoCoCoDa," *IEEE Access*, vol. 8, pp. 154 087–154 094, 2020.

[19] J. Han, K. Qian, M. Song, Z. Yang, Z. Ren, S. Liu, J. Liu, H. Zheng, W. Ji, T. Koike, X. Li, Z. Zhang, Y. Yamamoto, and B. W. Schuller, "An Early Study on Intelligent Analysis of Speech Under COVID-19: Severity, Sleep Quality, Fatigue, and Anxiety," in *Proc. of Interspeech*. Shanghai, China: ISCA, 2020, pp. 4946–4950.

[20] 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. M. Rothkrantz, J. Zwerts, J. Treep, and C. Kaandorp, "The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates," in *Proc. of Interspeech*. Brno, Czech Republic: ISCA, 2021, to appear.

[21] A. Muguli, L. Pinto, N. R., N. Sharma, P. Krishnan, P. K. Ghosh, R. Kumar, S. Ramoji, S. Bhat, S. R. Chetupalli, S. Ganapathy, and V. Nanda, "DiCOVA Challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics," 2021, arXiv.

[22] B. Rocha, L. Mendes, I. Chouvarda, P. Carvalho, and R.P.Paiva, "Detection of Cough and Adventitious Respiratory Sounds in Audio Recordings by Internal Sound Analysis," in *Proc. of the* 3^{rd} *International Conference in Biomedical and Health Informatics*. Thessaloniki, Greece: Springer, 2017, pp. 51–55.

[23] B. M. Rocha, L. Mendes, R. Couceiro, J. Henriques, P. Carvalho, and R. P. Paiva, "Detection of Explosive Cough Events in Audio Recordings by Internal Sound Analysis," in *Proc. of the* 39^{th} *Annual International Conference of the Engineering in Medicine and Biology Society*, Jeju, South Korea, 2017, pp. 2761–2764.

[24] X. Sun, Z. Lu, W. Hu, and G. Cao, "SymDetector: Detecting Sound-Related Respiratory Symptoms Using Smartphones," in *Proc. of the International Joint Conference on Pervasive and Ubiquitous Computing*. Osaka, Japan: ACM, 2015, pp. 97–108.

[25] B. McFee, M. McVicar, S. Balke, C. Thomé, C. Raffel, D. Lee, O. Nieto, E. Battenberg, D. Ellis, R. Yamamoto, J. Moore, R. Bittner, K. Choi, P. Friesch, F.-R. Stöter, V. Lostanlen, S. Kumar, S. Waloschek, S. Kranzler, R. Naktinis, D. Repetto, C. F. Hawthorne, C. Carr, W. Pimenta, P. Viktorin, P. Brossier, J. F. Santos, J. Wu, E. Peterson, and A. Holovaty, "librosa/librosa: 0.8.0," 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3955228

[26] L. Carpentier, D. Berckmans, A. Youssef, D. Berckmans, T. van Waterschoot, D. Johnston, N. Ferguson, B. Earley, I. Fontana, E. Tullo, M. Guarino, E. Vranken, and T. Norton, "Automatic cough detection for bovine respiratory disease in a calf house," *Biosystems Engineering*, vol. 173, pp. 45–56, 2018.

[27] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised Speech Activity Detection Using Voicing Measures and Perceptual Spectral Flux," *Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, 2013.

[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.

[29] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical Attention Networks for Document Classification," in *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, CA, USA: ACL, 2016, pp. 1480–1489.