ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Interpretability analysis of deep models for COVID-19 detection

Daniel Peixoto Pinto da Silva[1], Edresson Casanova[2], Lucas Rafael Stefanel Gris[3], Marcelo Matheus Gauy[4]*, Arnaldo Candido Junior[5], Marcelo Finger[4], Flaviane Romani Fernandes Svartman[6], Beatriz Raposo de Medeiros[7], Marcus Vinícius Moreira Martins[8], Sandra Maria Aluísio[2], Larissa Cristina Berti[9], and João Paulo Teixeira[10]

[1]Academic Department of Computing, Federal University of Technology – Paraná, Medianeira, Paraná, Brazil

[2]Department of Computer Science, Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, São Paulo, Brazil

[3]Institute of Informatics, Federal University of Goiás, Goiania, Goiás, Brazil

[4]Department of Computer Science, Institute of Mathematics and Statistics, University of São Paulo, São Paulo, São Paulo, Brazil

[5]Department of Computing and Statistics, Institute of Biosciences, Humanities and Exact Sciences, São Paulo State University, São José do Rio Preto, São Paulo, Brazil

[6]Department of Classical and Vernacular Literature, Faculty of Philosophy, Language, Literature and Human Sciences, University of São Paulo, São Paulo, São Paulo, Brazil

[7]Department of Linguistics, Faculty of Philosophy, Language, Literature and Human Sciences, University of São Paulo, São Paulo, São Paulo, Brazil

[8]Department of Literature and Linguistics, University of the State of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

[9]Department of Speech Therapy, Faculty of Philosophy and Sciences, São Paulo State University, Marília, São Paulo, Brazil

[10]Department of Eletronics, Research Centre in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança, Bragança, Portugal

***Corresponding author:**
Marcelo Matheus Gauy
(marcelo.gauy@usp.br)

Citation: da Silva DPP, Casanova E, Gris LRS, *et al*. Interpretability analysis of deep models for COVID-19 detection. *Artif Intell Health*. 2024;1(3):114-126. doi: 10.36922/aih.2992

Received: February 21, 2024

Accepted: June 17, 2024

Published Online: July 30, 2024

Abstract

During the coronavirus disease 2019 (COVID-19) pandemic, various research disciplines collaborated to address the impacts of severe acute respiratory syndrome coronavirus-2 infections. This paper presents an interpretability analysis of a convolutional neural network-based model designed for COVID-19 detection using audio data. We explore the input features that play a crucial role in the model's decision-making process, including spectrograms, fundamental frequency (F0), F0 standard deviation, sex, and age. Subsequently, we examine the model's decision patterns by generating heat maps to visualize its focus during the decision-making process. Emphasizing an explainable artificial intelligence approach, our findings demonstrate that the examined models can make unbiased decisions even in the presence of noise in training set audios, provided appropriate preprocessing steps are undertaken. Our top-performing model achieves a detection accuracy of 94.44%. Our analysis indicates that the analyzed models prioritize high-energy areas in spectrograms during the decision process, particularly focusing on high-energy regions associated with prosodic domains, while also effectively utilizing F0 for COVID-19 detection.

Keywords: Coronavirus disease 2019 detection; Voice processing; Gradient-weight class activation mapping

1. Introduction

In December 2019, a novel coronavirus, namely severe acute respiratory syndrome coronavirus-2, was identified as the causative agent for coronavirus disease 2019 (COVID-19). This coronavirus variant rapidly became a global concern, reaching pandemic status as declared by the World Health Organization.[1] COVID-19 evolved to become more contagious and lethal over a short period.

Researchers from all fields joined efforts to tackle the pandemic crisis. In particular, researchers in artificial intelligence (AI) and related areas sought methods to simplify COVID-19 detection. These methods use a variety of sources, such as medical examinations,[2] symptoms,[3] and X-ray images,[4] among others.[5] A potential source for COVID-19 detection is audio recordings. Several projects have collected audio samples from patients, including speech and cough sounds,[6-8] to develop detection models. These models could optimize the patient screening process. However, existing approaches have limitations in data collection procedures. For example, environmental noise can be present during audio capture, leading to model overfitting on such noise.

The dataset presented in the SPIRA project[8,9] illustrates these challenges. Positive audio samples (read speech) from COVID-19 patients were collected in hospitals, while samples from symptom-free individuals were obtained through a web interface. These samples were labeled as the control group, with the caveat that no additional testing for COVID-19 was performed on these subjects. Training a model on this dataset requires precautions to avoid learning biases due to differences in the collection environment, as patient audios may contain hospital noise, while the control group may include other environmental noises. Moreover, a model trained on this dataset contrasts healthy cases with more severe COVID-19 cases, which typically exhibit symptoms such as respiratory insufficiency. Such models will likely not be able to identify COVID-19 cases that do not induce severe symptoms.

In this work, we trained and analyzed convolutional neural networks (CNNs) for COVID-19 detection from audios using the dataset from the study by Casanova *et al*.[9] In addition, we analyzed factors important for the model decision using several criteria, namely spectrograms, fundamental frequency (F0), fundamental frequency standard deviation (F0-STD), speaker age, and sex. We applied the gradient-weight class activation mapping (Grad-CAM)[10] algorithm to generate heat maps, allowing us to investigate which pieces of information are most relevant for the model's classification decisions. As the dataset used in this work contained audios from different collection environments (hospital and domestic), learning biases could occur toward hospital noise. To mitigate this problem, we introduced hospital noise into domestic audio samples following the proposal of Casanova *et al*.[9] We also used their data augmentation techniques to improve model performance. Finally, we could literally hear the areas in the audio that the model values the most in its decision process. To achieve this, we multiplied the heat maps obtained from Grad-CAM by the original log-Mel spectrograms, and the result was synthesized. It is important to note that we focused on spectrograms rather than Mel-frequency cepstral coefficients (MFCCs)[11] to enhance interpretability, while previous works opted to explore MFCCs[9,12] to attain accuracies above or close to 90%. As spectrogram-based models were shown in those papers to have inherently lower accuracy, we employed methods such as transfer learning (e.g., pre-trained models on large-scale audio datasets) to recover the model's performance using log-Mel spectrograms.

As a result, our best model uses a pre-trained audio neural network (PANN)[13] called CNN14, which, through transfer learning, achieves 94.44% accuracy, in line with the accuracy on the same dataset from recent works[12] using transformers-based architectures.[14]

This work presents four main contributions:

(i) We present an analysis detailing the features crucial for deep models to detect or rule out COVID-19 in patient and control audios. In the analyzed data, spectrograms contain more discriminative information than the combination of F0, F0-STD, sex, and age. A visual analysis of heat maps generated by Grad-CAM shows that, among F0, F0-STD, sex, and age, the most important feature is F0.

(ii) We present an interpretation of the decisions made by deep models using heat maps and audio synthesis, following an explainable AI approach. Based on the heat maps and audio resynthesis, we formulate a few hypotheses for the factors affecting model decisions, such as (a) the structure of pauses (patients have longer and more frequent pauses than controls); (b) signal energy over time decreases faster for patients than controls; and (c) an interplay between syntax and prosody emerges as a boundary marked by formant vowel high energy.

(iii) Through manual analysis of the audio signals (using Grad-CAM), we ensure that the deep models focus on the voice (or silent pauses) rather than on environmental noise.

(iv) We demonstrate that models pre-trained on large-scale audio datasets, such as CNN14, can, through transfer learning, achieve accuracies on par with the best previously reported models,[12] even when using log-Mel spectrograms as input instead of MFCCs.

2. Related work

In the literature, COVID-19 detection has been studied using different types of input features for classification. From the perspective of feature analysis, these inputs can be roughly grouped into two categories: White-box or black-box, based on their ease of interpretation.

An example of an approach using mostly white-box features is the work of Bartl-Pokorny *et al.*[15] The authors used 88 features extracted from audios containing vowels to measure how COVID-19 patients differ from the control group. They found that F0-STD commonly varies between these two groups. In our work, we also used F0, F0-STD, and included sex and age as inputs for our deep models to detect COVID-19. Sex and age were included following the findings of previous works,[16-18] which identified that these factors influence F0 and F0-STD in COVID-19 patients. It was found that women and elderly subjects present more differences in these two parameters, as their voices become higher-pitched and less stable. Moreover, the study by Fernandes-Svartman *et al.*[18] demonstrated that the structure of pauses in speech undergoes significant changes between controls and hospitalized COVID-19 patients, even proposing a white-box model, which achieves above 87% accuracy using solely the speech pause distribution.

Regarding black-box features, Schuller *et al.*[19] proposed a challenge for COVID-19 detection from both speech and cough audios using the Cambridge COVID-19 Sound database.[6,7] They performed baseline experiments and identified thousands of features that can be used for general audio processing and, in particular, COVID-19 detection in audio. Zheng *et al.*[11] presented another example of black-box features, where MFCCs proved to be a useful method for COVID-19 detection while consuming few computational resources.[9] More robust approaches use spectrograms, transfer learning, and data augmentation for the task.[20] Recently, transformer-based architectures with MFCCs as input were used alongside transfer learning in the study by Gauy and Finger,[12] achieving accuracy above 95%. CNN-based PANNs (e.g., CNN14), which use spectrograms as input, were also used in a study by Gauy *et al.*,[21] achieving comparable accuracy to transformer-based architectures. Based on these results, we investigated the use of spectrograms for COVID-19 detection. In addition, transfer learning and data augmentation were employed in the study.

Related work either uses white-box features (e.g., sex and age) to better understand the effect of COVID-19 on patient's audio or black-box features (e.g., spectrograms) alongside deep learning for higher accuracy in COVID-19 detection tasks. In this work, we proposed an interpretability analysis of the decisions made by the deep learning models found in the literature, aiming for a better understanding of their results. To achieve this, we proposed the use of the Grad-CAM algorithm,[10] analyzing its heat maps and synthesizing audios based on those heat maps. These operations provide valuable insights into how deep models make their decisions. A similar approach to ours can be found in the study by Sobahi *et al.*,[22] where the authors used Grad-CAM to visualize the results generated by their proposed model for COVID-19 detection through cough sounds. Grad-CAM allowed them to identify which regions of the input were most relevant to the model's decision-making process. In addition, in a slightly different domain, previous works have used Grad-CAM to analyze COVID-19 detection models based on chest X-ray images.[23,24]

3. Methods

We used the SPIRA dataset from a previous study,[9] which contains spoken utterances from 432 speakers, including both patient and control group members. Audios were collected in COVID-19 wards where patients were hospitalized due to respiratory insufficiency, conventionally defined as a blood oxygen saturation level below 92%. Control group members were recorded using an application over the Internet. We used the same division into training, validation, and test sets as the previous study,[9] maintaining a balance by age and sex. Specifically, the dataset was divided into 292 training audios, 32 validation audios, and 108 test audios. The dataset includes recordings of patients and control group members speaking an utterance with no pre-defined pauses. The utterance is simple enough for most to understand but complex enough to present several polysyllabic words with primary and secondary stress syllables. The specific utterance was "*o amor ao próximo ajuda a enfrentar o coronavírus com a força que a gente precisa*" ("love for your neighbor helps face the coronavirus with the strength we need"). The dataset used is available at https://github.com/SPIRA-COVID19/SPIRA-ACL2021, and the codes for each model and experiment can also be found at https://github.com/danpeixoto/covid19-interpretability-analysis.

In this work, inspired by previous approaches,[12,13] we employed transfer learning methods from pre-trained models (described in Section 3.1). Following established methods,[20] we explored data augmentation techniques (described in Section 3.2). Similar to previous studies,[9,12] we utilized audio splitting based on windowing (described in Section 3.3). During training, we performed preprocessing steps (described in Section 3.4) on the audios, termed dynamic preprocessing,[9] to tackle overfitting issues. By combining all the aforementioned techniques, we performed six experiments, described in

Sections 3.5 and 3.6. The goal of these experiments was to determine which operations are relevant for classification and how they affect the model's decision process (analyzed by Grad-CAM).

The server used for our training has an Intel Xeon Silver CPU processor (39 cores, 2.40 GHz), 56 GB of RAM, and two Nvidia 2080 GPUs (8 GB of VRAM each). Some of the runs occurred only in the CPU cores, while others used both GPUs and the CPU. Overall, all the experiments took approximately the same time to run, around 6 h in the CPU-only scenario or 1 h using both GPUs and CPU. Some small variations were observed, mainly due to the preprocessing techniques used, as they were performed exclusively on the CPU during each epoch of training. It should be noted that inference took only a few seconds in our environment.

3.1. Transfer learning with PANNs

PANNs have proven effective for transfer learning across various tasks.[13] They have been successfully applied to multiple audio classification tasks, such as audio set tagging,[13] speech emotion recognition,[25] and automated audio captioning.[26] PANNs are Mel spectrogram-based models and trained on the AudioSet dataset, which comprises approximately 1.9 million audios, totalizing 527 classes and over 5000 h. While the original authors explored several architectures, in this work, we used only the CNN14 architecture due to its simplicity and similarity to SpiraNet,[9] allowing it to benefit from the same preprocessing techniques.

3.2. Data augmentation

Following the work of Casanova *et al.*,[20] three data augmentation techniques were applied: Noise insertion, Mix-up, and SpecAugment.

First, noise insertion was performed due to the different recording environments present in the SPIRA dataset for patient and control group audios. Previous research[9] has shown that models trained on this dataset can overfit if the data are not preprocessed adequately, leading to biases, such as distinguishing control and patient groups based on the presence of hospital ward noise. To mitigate this, we followed the Casanova *et al.*[9] approach by injecting hospital ward noise into all audios. For some experiments, we inserted four noise recordings for the control group and three for the patient group, while other experiments used three audio recordings for each class, based on Casanova *et al.*'s findings[9].

Second, due to the small size of the training set, we used a data augmentation technique called Mix-up to increase model robustness. Mix-up combines two random instances (*xi* and *xj*) from the training set and their respective classes (*yi* and *yj*) to generate a new instance[27] $\left(\tilde{x}, \tilde{y}\right)$ using Equations I and II:

$$\tilde{x} = \lambda x_i + \left(1-\lambda\right) x_j \quad \text{(I)}$$

$$\tilde{y} = \lambda y_i + \left(1-\lambda\right) y_j \quad \text{(II)}$$

where $\lambda \in [0, 1]$ is generated from a Beta distribution.

Unlike common image processing augmentation techniques, such as rotation, cropping, and horizontal flipping, Mix-up is applicable to various tasks, including audio processing.[28] It helps generate better decision frontiers in the manifold extracted by the model during training, which is particularly beneficial for small datasets.

Finally, to further enhance model robustness in the cases of small training sets, we used the SpecAugment[29] data augmentation technique. SpecAugment is designed for spectrograms and was initially developed for automatic voice recognition. It performs augmentation on the spectrogram by first applying distortion in the time dimension, termed time warping in the study by Casanova *et al.*,[9] and then masking parts of frequency channels and masking blocks in time. The frequency mask is applied over f consecutive Mel channels $[f0, f0 + f)$, where f is chosen uniformly from 0 to F and F represents the maximum number of masked Mel channels (set to eight in our experiments). The parameter $f0$ is uniformly chosen at random from $[0, v - f)$, where v is the total number of Mel-frequency channels. The temporal mask is performed over t time slots $[t0, t0 + t)$, where t and $t0$ are determined analogously to the frequency mask.

3.3. Windowing

Each original audio, which is at least 4 s long, is divided into smaller 4-s audios. The division was performed using a 4-s window with a 1-s hop. For example, a 5-s audio was split into two segments: The first from seconds 0 – 4 and the second from seconds 1 – 5. This approach, initially employed by Casanova *et al.*,[9] ensures uniform audio lengths and prevents model overfitting based on audio lengths. As patient audios tend to be longer, models can overfit on audio length if no normalization is done.

The windows cover repeated fragments of the original audios to include as many fragments of the original spoken sentence as possible. It is important to note that windowing was performed separately for training and test sets. During training, each fragment was labeled with the class of the original audio. In the test set, a voting mechanism over the windowed audios was used to determine the class

of original audio, as described by Casanova *et al.*[9] The voting summed the predicted probabilities for each class. Windowing also served as a simple data augmentation technique, in addition to the approaches presented in Section 3.2.

3.4. Dynamic preprocessing

The audios were preprocessed for each training step, ensuring a richer variety of augmented data. To maintain our model consistent, the same preprocessing was applied during the validation and test phases. The following operations were carried out:

(i) Noise injection
(ii) Windowing
(iii) Spectrogram extraction
(iv) Spec-augment application (only for training)
(v) Mix-up application (only for training)
(vi) Training step/test step.

Operations 4 and 5 were applied only to PANN-based experiments and only during training, while the other operations were common to all experiments. For operation 3, we used different parameters for spectrogram extraction in our experiments. Table 1 presents the two settings used across the experiments presented in Sections 3.5 and 3.6: Set 1 was used for SpiraNet and matched the parameters from Casanova *et al.*[9] and Set 2 was used for CNN4 and needed to be consistent with the parameters used in pre-training. Two parameters were common for all spectrogram-based experiments: The number of fast Fourier fransform[30] components (1200) and the spectrogram format (log-Mel).

3.5. Experiments to find the best inputs

Here, we describe three experiments aiming to estimate the accuracy of the SpiraNet[9] with respect to three different input configurations. These experiments investigated the role of different information types (spectrogram, F0, F0-STD, age, and sex) in the model's decision process. Spectrograms are matrices, while F0 is a vector, and the remaining data are scalars. We converted all these data into matrices to facilitate visual analysis using Grad-CAM, described in the subsequent sections. The representation is shown in Figure 1. The input, in its full form, has 401 × 120 pixels, where the spectrogram occupies the top region (401 × 80). Age, F0-STD, and sex occupy 20 lines, while age and sex use 133 columns and F0-STD uses 135 columns. Age is represented by shades of gray, as it is a scalar value, and F0-STD is similarly represented. Sex is a binary value, with zero for males and one for females. F0 is represented in a "bar code" style, where each value in the original vector is repeated across an entire column in the generated matrix.

Table 1. Settings used in the experiments

Set	Hop size (ms)	Number of frequency	Number of Mel	Window length (ms)
1	160	601	80	400
2	320	513	64	1,024

Abbreviation: ms: Milliseconds.

Using the scheme presented in Figure 1, the first three proposed experiments are:

- Experiment 1: Uses only the spectrogram (401 × 80 pixels) as input
- Experiment 2: Uses F0, F0-STD, age, and sex (401 × 40 pixels) as input
- Experiment 3: Uses all input data present in Figure 1, including the spectrograms, F0, F0-STD, age, and sex (401 × 120 pixels).

All three experiments are based on the SpiraNet model and use the configurations from Set 1 of Table 1. Moreover, the general hyperparameters for all the experiments (including Experiments 4 and 5 in Section 3.6), based on Casanova *et al.*,[9] are as follows: Binary cross-entropy loss and the Adam optimizer.[31] Given that the focus is on studying the model's decision process rather than performance, the batch size is set at one, early stopping and a learning rate scheduler are not used, and the number of epochs is set to 1000 for all experiments. Despite these settings, CNN14 achieves accuracies close to the best models reported in the literature. We used a fixed learning rate of 0.001 and a weight decay of 0.01.

3.6. Experiments over the training process

We performed three additional experiments to analyze classification models with respect to potential changes during training, pre-training, and post-processing. The three experiments are described as follows:

- Experiment 4: The goal of this experiment is to determine how the accuracy of a classification model changes when using large-scale pre-trained models. To achieve this, it focuses on pre-training, exploring the use of transfer learning through a PANN model (CNN14). This experiment was configured using Set 2 from Table 1.

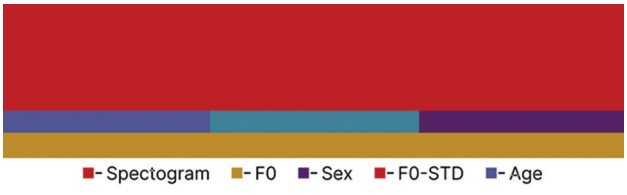

Figure 1. Input representation. Notes: F0: Fundamental frequency; F0-STD: Fundamental frequency standard deviation

- Experiment 5: This experiment aims to explore how the accuracy of a classification model changes when data augmentation techniques, such as SpecAugment and Mix-up, are used. In this experiment, we used the SpiraNet model but replaced MFCCs with spectrograms to simplify human analysis and improve audio resynthesis (see Experiment 6). As usual, for the SpiraNet, we use Set 1 of Table 1.
- Experiments 6a and 6b: These experiments differ from the previous five. We performed a qualitative analysis focused on model explainability using heat maps generated by Grad-CAM. This method aims to uncover the underlying reasons for the model's classification decisions by generating heat maps that highlight important zones in the decision process. First, in Experiment 6[a], we conducted a preliminary analysis and case study, investigating Grad-CAM from Experiments 1 to 3 (see Section 4.2) to understand which parts of the input are more relevant for classification. Then, in Experiment 6b, we performed a detailed analysis, focusing on the heat maps generated in Experiment 1. Our preliminary analysis showed that the spectrogram plays a major role in classification (see Section 4.3). In Section 4.3, we also resynthesized audios from Experiment 1, allowing us to hear them and investigate attention from both a visual and aural perspective. The audio reconstruction process is done in two steps. First, the heat map generated by Grad-CAM and the log-Mel spectrogram are combined using the Hadamard product. Second, the result and the phase of the original spectrogram are used to generate new audios highlighting the moments and frequencies the model considered most important in its decisions. We refer to the combination of original log-Mel spectrograms with heat maps as modified spectrograms.

4. Results

4.1. Experiments 1 – 5: Quantitative analysis

Table 2 presents the results of Experiments 1 – 5, with accuracies ranging from 65.74% to 94.44%. From Experiment 1, we observe that spectrograms are discriminative. Likewise, Experiment 2 showed that F0, F0-STD, sex, and age also contain discriminative information. However, spectrograms appear to carry more useful information since the accuracy of Experiment 1 is >10% higher than that of Experiment 2. Experiment 3 suggests that features extracted from inputs in Experiments 1 and 2 are largely equivalent despite a slight increase in accuracy (almost 2%) compared to Experiment 1. It should be noted that Experiments 1 - 4 used only noise insertion as data augmentation.

Table 2. Results from Experiments 1 – 5

Experiment	True		False		Accuracy (%)
	Positives	Negatives	Positives	Negatives	
1	37	49	5	17	79.63
2	36	38	16	18	68.52
3	44	44	10	10	81.48
4	51	51	3	3	94.44[a]
5	49	22	32	5	65.74

Note: [a]The highest accuracy among the experiments was achieved by Experiment 4.

Experiment 4 achieved the highest accuracy (94.44%), indicating that transfer learning significantly impacts learning features from patient and control groups, surpassing the results of Casanova *et al.*[9] These findings suggest that CNN14 might be better suited than SpiraNet for COVID-19 detection. CNN14's results are comparable in accuracy to those of transformers-based architectures described by Gauy and Finger,[12] with the added advantage of using spectrograms as input instead of MFCCs, as was the case for the MFCC-transformer.[12] This advantage is attributed to the effectiveness of the transfer learning used. Experiment 5 demonstrates that data augmentation techniques (SpecAugment and Mix-up) did not improve SpiraNet accuracy, as it performed worse than in Experiments 1 and 2. Experiments 6a and 6b are presented separately because they are based on heat maps, human analysis, and audio resynthesis (Sections 4.2 and 4.3, respectively).

Regarding errors, most experiments resulted in a balance of false positives and false negatives. Experiment 1 was an exception, presenting more false negatives. This experiment might have been more susceptible than others to cases of silent hypoxia, in which a patient has low blood oxygenation but does not present severe symptoms. Another exception was Experiment 5, which had significantly more false positives (32) than false negatives (5). A hypothesis for this phenomenon is that SpecAugment forces the model to give less importance to pauses, which are crucial for detecting respiratory insufficiency.[18] This may occur because the method introduces artificial pauses in training data.

4.2. Experiment 6a: Case study based on heat map analysis

Experiment 6a involved using Grad-CAM to generate heat maps for experiments based on inputs (Section 3.5). Figures 2-4 present the results of heat maps and modified spectrograms for Experiments 1 – 3, respectively.

Experiment 1 focused solely on spectrograms. The visual results for two patients and two control group

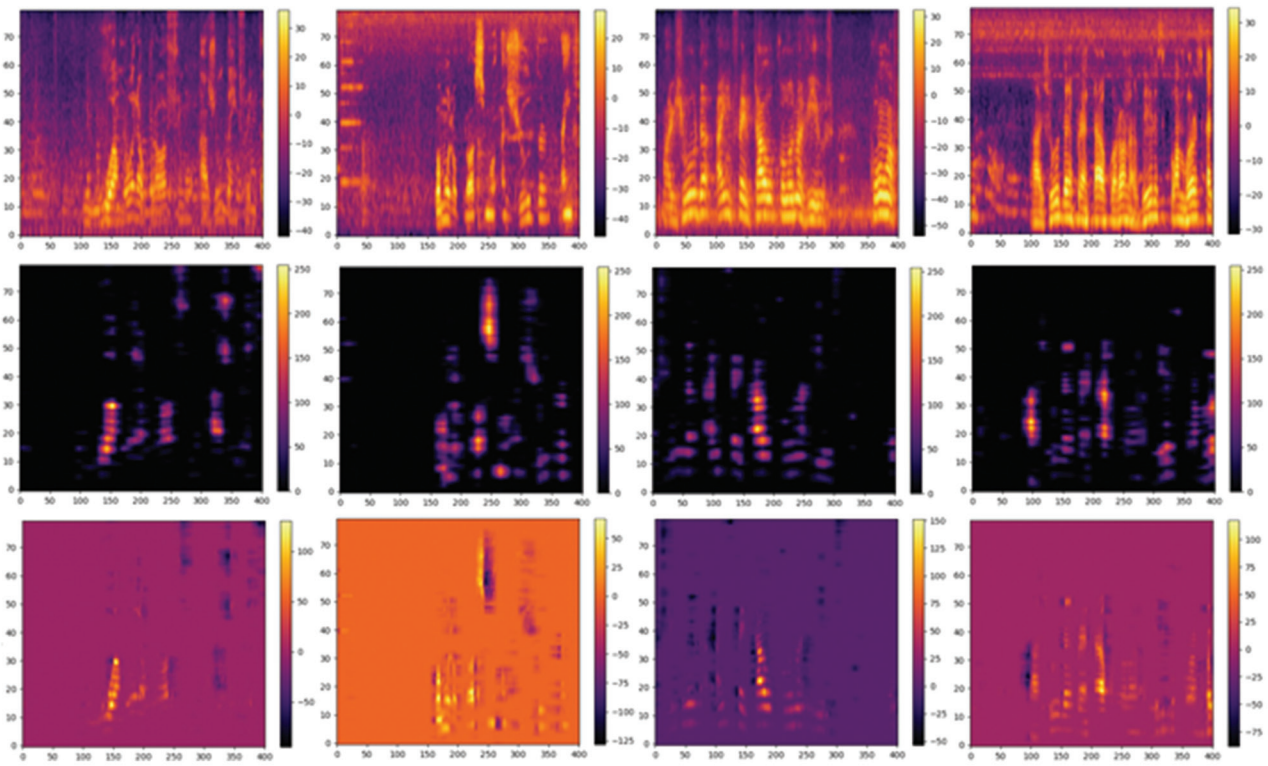

Figure 2. Results from Experiment 6a regarding Experiment 1 (spectrogram only), including original spectrograms (top), heat maps (middle), and modified spectrograms (bottom) for two control group members (left) and two patients (right).

members, including original spectrograms, heat maps, and modified spectrograms, are presented in Figure 2. We observe activity (attention) in high-energy regions over the input. These results suggest that energy levels and audio formats (*H*2 and *H*3) may play a significant role in COVID-19 detection.

In Section 4.1, the results indicated F0, F0-STD, age, and sex as distinctive features for COVID-19 detection. Figure 3 presents Experiment 2 visual representations for two patients and two control group members. It can be noted that F0 plays a major role in this model's detection process, especially in regions associated with transitions from voiced phonemes to pauses (*H*1) or to voiceless phonemes. The same applies to transitions from pauses or voiceless phonemes to voiced phonemes. In addition, sex and age appear to play a role in control classification, although not as noticeable as F0. On the other hand, F0-STD appears to be disregarded by the model.

Figure 4 presents the visual representations generated using all available information (spectrograms, F0, F0-STD, sex, and age). Heat maps suggest that spectrograms, F0, and sex are useful for patient classification, while control group detections are based only on spectrograms and F0. These observations indicate that spectrograms and F0 may contain complementary information, given the slightly superior accuracy obtained from Experiment 3 compared to Experiments 1 and 2.

4.3. Experiment 6b: Phonetical investigation and qualitative analysis

The phonetic investigation and qualitative analysis presented here were carried out by three linguists. Four main inputs were considered:

(i) Regular spectrograms in hertz were obtained from the original audios. These spectrograms were generated using the software PRAAT v6.1.09.
(ii) Original and modified Mel-spectrograms to highlight attention, as presented in Section 4.2 (Experiment 1).
(iii) Resynthesized audios from the modified spectrograms from the previous input allow us to hear where the model pays attention, while spectrograms show where the model focuses. These resynthesized audios are publicly available (https://drive.google.com/drive/folders/1aQEq82iUpnAmrQzQ52458GORv8PEK3nr?usp=share_link).
(iv) Regular spectrograms in hertz from audios resynthesized from our modified spectrograms obtained from the previous input. These spectrograms combine speech with heat maps.

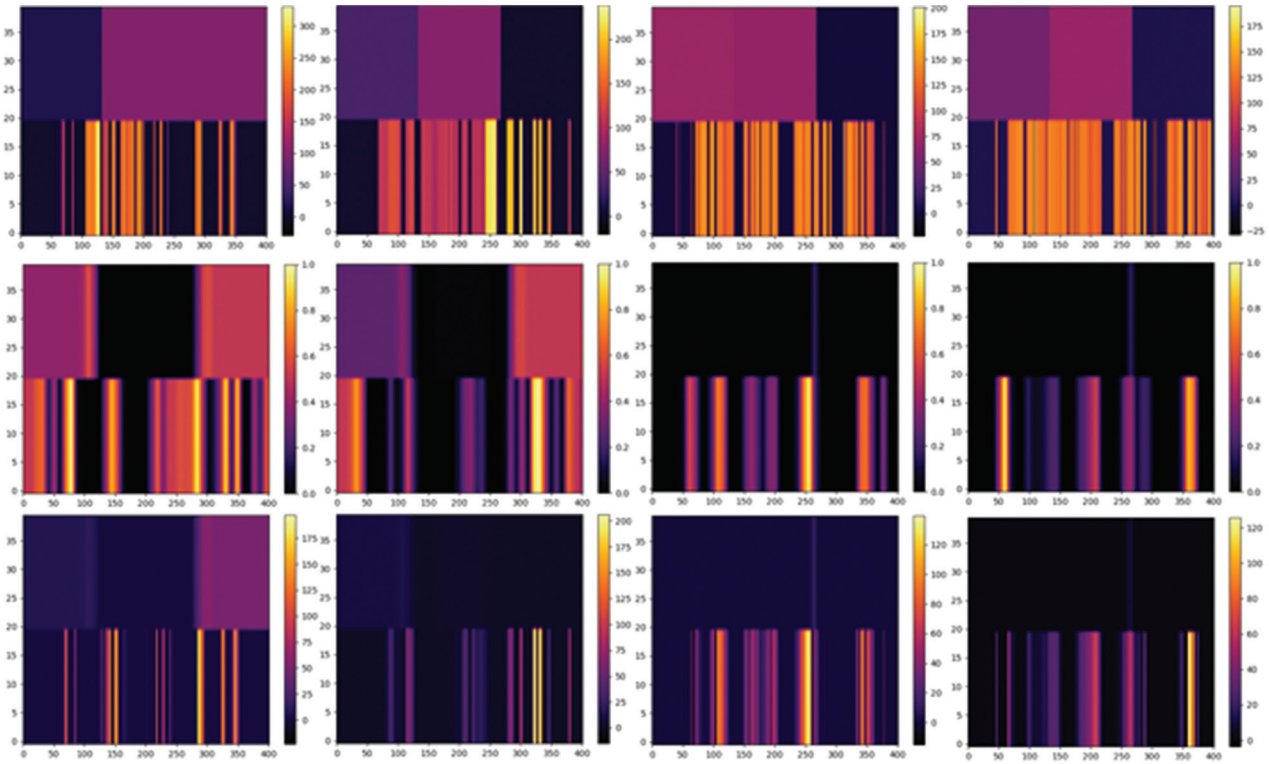

Figure 3. Results from Experiment 6a regarding Experiment 2 (all inputs except spectrograms), including original images (top), heat maps (middle), and modified images (bottom) for two control group members (left) and two patients (right).

Given our windowing approach, which involved generating 4-s windows with a 1-s hop, this analysis considered only the central audio fragment in the window. In total, the central fragments of 73 audios were inspected, containing 30 correct predictions for each class and 13 errors, from seven patients and six controls. It is important to note that Experiment 6b resynthesis is based on the model trained in Experiment 1. We chose Experiment 1 rather than 4 for better comparability with the analysis performed in Section 4.2.

It was observed that the decision process usually hinges on two aspects of the speech sound signal: The continuity of the signal versus its interruption. Thus, the model appears to pay attention to an alternation between the continuity of speech sounds and their discontinuity, which, in terms of intonational analysis, are pauses inserted by speakers. This observation is in line with the findings of Fernandes-Svartman *et al.*,[18] which noted that patients' pauses are significantly longer than those of control subjects and, being more frequent, are inserted in more places throughout the utterance.

Considering short-term parameters, such as the most salient vowels for the model, it was observed that the vowels/a/from "*ajuda a*" ("helps to") and "*enfrentar*" ("to face");/ɔ/from "*próximo*" ("neighbor");/o/from "*força*" ("strength");/oN/from "*com*" ("with") are those reproduced with more intensity in the modified spectrogram. This pattern corresponds to what occurs in the original audio spectrogram. Besides being expected, it is reasonable, since these vowels occupy prominent places in the utterance or are intrinsically more intense, such as the low vowels/a/and/ɔ/. On the other hand, the mid-high, oral/o/, and nasal/oN/ vowels do not have the same sound amplitude as the low ones but occupy a prosodically highlighted place in the utterance.

Therefore, on the one hand, we have the phonetic features of vowels and, on the other hand, the prosodic feature interacting with morpho-syntactic and semantic-pragmatic levels. The interaction discussed here explains the emphasis on the verb (there is usually a peak of F0 in the verbal item in a statement). In semantic terms, emphasis is given to "*com a força*" ("with the strength"). The initial expression of the adverbial phrase "*com a força que a gente precisa*" ("with the strength we need") is often phrased as an intonational phrase in our data. The intonational phrase initial position is prominent in Portuguese.[32,33] In pragmatic terms, this adverbial phrase is new information that modalizes the meaning of the verb "to face" (it is necessary to face the virus with strength).

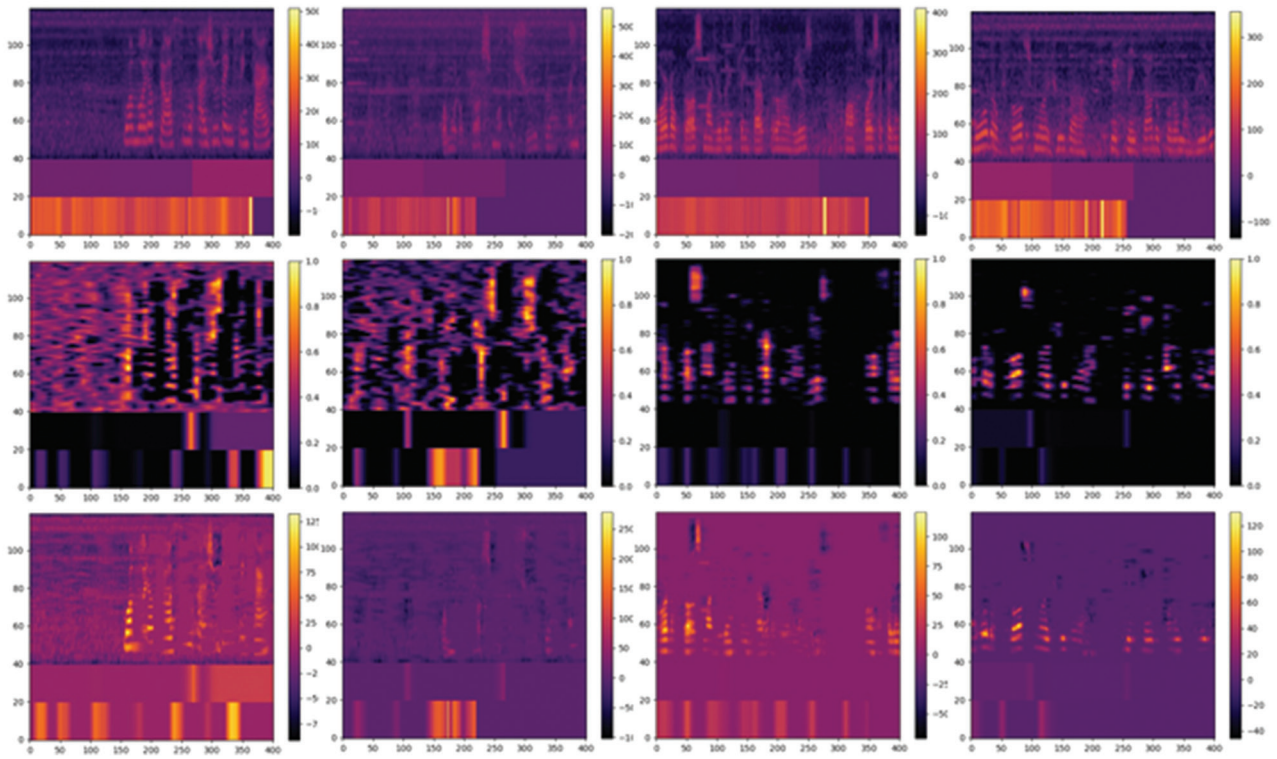

Figure 4. Results from Experiment 6a regarding Experiment 3 (all inputs), including original images (top), heat maps (middle), and modified images (bottom) for two control group members (left) and two patients (right).

Taking into account the phonetic analysis in interaction with other linguistic levels, the model shows the following preferences for distinguishing patients from controls:

i. Average patient pauses (approximately 400 ms) represent large interruptions of formant frequency tracks.
ii. Intrinsically more intense vowels are important clues. The model highlighted that the first (F1) and second (F2) formants in a 550 – 1300 Hertz range for both groups. For patients, these highlights can occur before or after a pause.
iii. The interaction between non-low vowels (with F1 and F2 ranging from 350 to 1000 Hz), the morpho-syntactic context, the prosodic domain in which they are produced, and their semantic-pragmatic role indicates that these segments are more prominent in the utterance, which appears important for the model.
iv. The interaction between non-low vowels and the initial position in the intonational phrase ("*com a força que a gente precisa*" ["with the strength we need"]) results in prosodic emphasis of this unit ("*com a força*" ["with the strength"]), which draws the model's attention.

Finally, for the correct predictions, regarding the interplay between the speech sound signal and the prosodic behavior that emphasizes it, we propose the following explanatory hypothesis: The model analyzes the signal as continuous and emphasizes some vowel formants in one of the groups; in another group, it focuses on important interruptions in a similar speech signal (the same sentence uttered by patients and controls). This approach leads to successfully distinguishing the two different speech groups, as patients with respiratory difficulties are unable to produce fluent speech and usually speak linguistic utterances with many pauses.

5. Discussion

In this section, we discuss a few hypotheses that can be deduced from our results (Section 5.1) as well as a few limitations of our approach and potential future work (Section 5.2).

5.1. Hypotheses for the model decision process

Regarding the question of which input features are best for the models, our results demonstrated that spectrograms convey important features for classification compared to other information, such as sex, gender, and F0-STD (Table 2). F0 also presented a small improvement during the classification process.

Regarding the training process, we found that noise insertion is important, consistent with previous findings;[20] therefore, we used it in all experiments. Other augmentations, such as Mix-up and SpecAugment, did not lead to improvements in the model. On the contrary, accuracy decreased. Transfer learning, on the other hand, proved to be important in this domain, as CNN14 achieved superior results compared to all other models and is comparable to the current state of the art in the literature for this task.

Furthermore, with respect to the training process, we noted in preliminary experiments some variance in the aspects a model can focus on during inference. The structure of pauses, syntactic boundaries, and pretonic syllables, among other factors, may be more or less evidenced by the models after training. This result is expected because artificial neural networks are high-variance, low-bias classifiers with randomized parameter initialization. We observed that transfer learning appears to reduce this variance.

Regarding the qualitative analysis, our first case study indicated that detailed evaluation would be better performed in the spectrograms-only scenario, which allowed for audio resynthesis, improving the process. As a result of this analysis, we can formulate the following hypotheses to explain the obtained variance and understand the data aspects that may play a role in model learning:

(i) H1: Pauses are important clues for detecting COVID-19 since patients tend to make more pauses for breathing than the control group.
(ii) H2: As the air starts decreasing in the lungs, the speaker may begin to lose breath, or the signal energy may begin to decrease. Thus, energy over time can be an important clue.
(iii) H3: An interplay between syntax and prosody is expected to emerge as a boundary marked by formant vowel high energy, i.e., phonetically.

The first hypothesis confirms that deep models use the discrepancy in the structure of pauses between patients and controls, as observed by Fernandes-Svartman *et al.*[18] The second and third hypotheses are newly observed discrepancies, which were found to be present by deep learning models.

Our work also confirms the hypothesis from previous works[9,12] that the addition of hospital ward noise, alongside suitable preprocessing steps, prevents the models from making biased decisions in the COVID-19 detection task. Through Grad-CAM analysis, we confirm that deep models focus on the voice (or silent pauses) rather than on environmental noise.

Finally, our best model (CNN14) achieved an accuracy of 94.44%. This number is almost as good as the best models reported in the literature[12] and shows that proper use of transfer learning can make log-Mel spectrogram input nearly as efficient as MFCC input.

5.2. Limitations and future work

In future works, we plan to investigate other audio-related features, such as autocorrelation, jitter, and shimmer. We also intend to investigate the beginning of a sentence. When a speaker starts to produce a sentence, they have more air in their lungs, which decreases as they speak. Some models may focus more on the audio at the beginning, measuring the signal energy, as the initial energy in the audio may provide hints about pulmonary capacity. In addition, we plan to investigate models of related diseases, such as general cases of respiratory insufficiency. Finally, we aim to investigate the variance in model training, identifying factors that are important for model inference and techniques that reduce variance in the learned models (such as transfer learning).

6. Conclusion

This work presents a method for interpretability analysis of audio classification for COVID-19 detection based on CNNs. Our work focuses on explainable AI. We investigated the importance of different features in the training process and generated heat maps to understand the model's reasoning for its predictions.

Regarding the input data, our results show that spectrograms are a suitable representation for COVID-19 detection. F0 appears to be almost as efficient as spectrograms, and the combination of these two inputs led to a small increase in the model performance. Grad-CAM analysis indicates that F0 is a more important feature than F0-STD, sex, and age. Moreover, Grad-CAM and audio resynthesis helped us formulate hypotheses about the factors that determine the model's classification process and confirm that the deep models used do not rely on environmental noise for decision-making. Our best model (CNN14) achieved 94.44% accuracy, on par with the best models in the literature.[12]

Acknowledgments

We gratefully acknowledge the support of NVIDIA corporation with the donation of a GPU used in part of the experiments presented in this research.

Funding

This work was supported by FAPESP grants 2022/16374-6 (MMG), 2020/06443-5 (SPIRA), and 2023/00488-5

(SPIRA-BM) and by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Arnaldo Candido Junior, Marcelo Finger
Investigation: Daniel Peixoto Pinto da Silva, Edresson Casanova, Arnaldo Candido Junior
Methodology: Daniel Peixoto Pinto da Silva, Lucas Rafael Stefanel Gris, Flaviane Romani Fernandes Svartman, Beatriz Raposo de Medeiros, Marcus Vinícius Moreira Martins, Larissa Cristina Berti
Writing – original draft: Daniel Peixoto Pinto da Silva, Arnaldo Candido Junior, Flaviane Romani Fernandes Svartman, Beatriz Raposo de Medeiros, Marcus Vinícius Moreira Martins, Larissa Cristina Berti
Writing – review & editing: Marcelo Matheus Gauy, Arnaldo Candido Junior, Sandra Maria Aluísio, João Paulo Teixeira, Marcelo Finger

Ethics approval and consent to participate

The research described in the paper was developed within the scope of the SPIRA Project (System for the Early Detection of Respiratory Insufficiency via Audio), which was approved by the Research Ethics Committee (IRB) of the Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HCFM/USP), Report 3.988.088, approved on April 24, 2020. The report states that this research does not require signed informed consent, as data collection involves voice assessment, and participants consent to participate by recording their acceptance on the equipment (cell phone) used in the study.

Consent for publication

Due to the pandemic, the IRB of the Hospital das Clínicas authorized us to collect patients' agreement to participate in the form of a recorded acceptance only. All participants expressed their agreement in a recorded audio.

Availability of data

The audio data can be found at https://github.com/SPIRA-COVID19/SPIRA-ACL2021/tree/master

Further disclosure

This paper has been uploaded to Arxiv at: https://arxiv.org/pdf/2211.14372.pdf. The code for the models can be found at: https://github.com/danpeixoto/covid19-interpretability-analysis.

References

1. *WHO Director-General's Opening Remarks at the Media Briefing on COVID-19*. World Health Organization; 2020. Available from: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19-11-march-2020 [Last accessed on 2024 Jul 19].

2. Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, Cabitza F. Detection of COVID-19 infection from routine blood exams with machine learning: A feasibility study. *J Med Syst*. 2020;44(8):135.

   doi: 10.1007/s10916-020-01597-4

3. Zoabi Y, Deri-Rozov S, Shomron N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *NPJ Digit Med*. 2021;4(1):3.

   doi: 10.1038/s41746-020-00372-6

4. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Acharya UR. Automated detection of COVID-19 cases using deep neural networks with x-ray images. *Comput Biol Med*. 2020;121:103792.

   doi: 10.1016/j.compbiomed.2020.103792

5. Acar E, Şahin E, Yılmaz İ. Improving effectiveness of different deep learning-based models for detecting COVID-19 from computed tomography (CT) images. *Neural Comput Appl*. 2021;33:17589-17609.

   doi: 10.1007/s00521-021-06344-5

6. Han J, Brown C, Chauhan J, *et al*. Exploring Automatic COVID-19 Diagnosis via Voice and Symptoms from Crowdsourced Data. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE; 2021:8328-8332.

   doi: 10.1109/icassp39728.2021.9414576

7. Brown C, Chauhan J, Grammenos A, *et al*. Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM; 2020:3474-3484.

   doi: 10.1145/3394486.3412865

8. Aluísio SM, Camargo Neto AC, Casanova E, *et al*. Detecting Respiratory Insufficiency via Voice Analysis: The SPIRA Project. In: *Practical Machine Learning for Developing Countries on the Tenth International Conference on Learning Representations*; 2022.

9. Casanova E, Gris L, Camargo A, *et al*. Deep learning against COVID-19: Respiratory insufficiency detection in Brazilian Portuguese speech. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP*; 2021:625-633.

   doi: 10.18653/v1/2021.findings-acl.55

10. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D,

Batra D. Grad-cam: Visual Explanations from Deep Networks Via Gradient-Based Localization. In: *Proceedings of the IEEE International Conference on Computer Vision*; 2017:618-626.

doi: 10.1109/ICCV.2017.74

11. Zheng F, Zhang G, Song Z. Comparison of different implementations of MFCC. *J Comput Sci Technol*. 2001;16(6):582-589.

doi: 10.1007/BF02943243

12. Gauy MM, Finger M. Audio MFCC-Gram Transformers for Respiratory Insufficiency Detection in COVID-19. In: *Proceedings XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana,* STIL; 2021:143-152.

doi: 10.5753/stil.2021.17793

13. Kong Q, Cao Y, Iqbal T, Wang Y, Wang W, Plumbley MD. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. Vol. 28. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing*; 2020:2880-2894.

doi: 10.1109/TASLP.2020.3030497

14. Vaswani A, Shazeer NM, Parmar N, *et al.* Attention is all you need. In: *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017:6000-6010.

15. Bartl-Pokorny KD, Pokorny FB, Batliner A, *et al.* The voice of COVID-19: Acoustic correlates of infection in sustained vowels. *J Acoust Soc Am*. 2021;149(6):4377-4383.

doi: 10.1121/10.0005194

16. Berti LC, Spazzapan EA, Pereira PL, *et al*. Mudanças Nos Parâmetros Acústicos da voz em Brasileiros com COVID-19 [Changes in the Acoustic Parameters of the Voice in Brazilians with COVID-19]. In: *XXIX Congresso Brasileiro e o IX Congresso Internacional de Fonoaudiologia*; 2021:2819-2819. [In Portuguese]

17. Berti LC, Spazzapan EA, Queiroz M, *et al*. Fundamental frequency related parameters in Brazilians with COVID-19. *J Acoust Soc Am*. 2023;153:576-585.

doi: 10.1121/10.0016848

18. Fernandes-Svartman F, Berti L, Martins M, Medeiros BR, Queiroz M. Temporal prosodic cues for COVID-19 in Brazilian Portuguese speakers. In: *Speech Prosody* 2022. ISCA; 2022:210-214.

doi: 10.21437/speechprosody.2022-43

19. Schuller BW, Batliner A, Bergler C, *et al*. The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates. In: *Interspeech* 2021. ISCA; 2021:431-435.

doi: 10.21437/interspeech.2021-19

20. Casanova E, Candido Jr. A, Fernandes Jr. RC, *et al.* Transfer Learning and Data Augmentation Techniques to the COVID-19 Identification Tasks in ComParE 2021. In: *Interspeech* 2021. ISCA; 2021:446-450.

doi: 10.21437/interspeech.2021-1798

21. Gauy MM, Berti LC, Cândido Júnior A, *et al*. Discriminant Audio Properties in Deep Learning Based Respiratory Insufficiency Detection in Brazilian Portuguese. In: *Artificial Intelligence in Medicine: 21st International Conference on Artificial Intelligence in Medicine*; 2023:271-275.

doi: 10.1007/978-3-031-34344-5_32

22. Sobahi N, Atila O, Deniz E, Sengur A, Acharya UR. Explainable COVID-19 detection using fractal dimension and vision Transformer with Grad-CAM on cough sounds. *Biocybern Biomed Eng.* 2022;42(3):1066-1080.

doi: 10.1016/j.bbe.2022.08.005

23. Moujahid H, Cherradi B, Al-Sarem M, *et al*. Combining CNN and Grad-CAM for COVID-19 disease prediction and visual explanation. *Intell Autom Soft Comput*. 2022;32(2):723-745.

doi: 10.32604/iasc.2022.022179

24. Panwar H, Gupta P, Siddiqui MK, Morales-Menendez R, Bhardwaj P, Singh V. A deep learning and Grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest x-ray and ct-scan images. *Chaos Solitons Fractals*. 2020;140:110190.

doi: 10.1016/j.chaos.2020.110190

25. Gauy MM, Finger M. Pretrained Audio Neural Networks for Speech Emotion Recognition in Portuguese. In: *First Workshop on Automatic Speech Recognition for Spontaneous and Prepared Speech Speech emotion recognition in Portuguese, SE&R*; 2022.

26. Xu X, Dinkel H, Wu M, Xie Z, Yu K. Investigating Local and Global Information for Automated Audio Captioning with Transfer Learning. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; 2021:905-909.

doi: 10.1109/ICASSP39728.2021.9413982

27. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: Beyond Empirical Risk Minimization. In: *International Conference on Learning Representations*; 2018.

28. Xu K, Feng D, Mi H, *et al.* Mixup-based acoustic scene classification using multi-channel convolutional neural network. In: Advances in Multimedia Information Processing. Vol. 11166. Cham: Springer; 2018. p. 14-23.

doi: 10.1007/978-3-030-00764-5_2

29. Park DS, Chan W, Zhang Y, *et al.* SpecAugment: A simple data augmentation method for automatic speech recognition. In: *Interspeech* 2019. ISCA; 2019:2613-2617.

doi: 10.21437/interspeech.2019-2680

30. Brigham EO, Morrow R. The fast fourier transform. *IEEE Spectrum*. 1967;4(12):63-70.

doi: 10.1109/MSPEC.1967.5217220

31. Kingma DP. Adam: A Method for Stochastic Optimization. In: *International Conference on Learning Representations*; 2014.

32. Frota S. *Prosody and Focus in European Portuguese: Phonological Phrasing and Intonation*. London: Routledge; 2014.

33. Tenani L. Domínios Prosódicos no Português do Brasil: Implicações Para Prosódia e Para a Aplicação de Processos Fonológicos [Prosodic Domains in Brazilian Portuguese: Implications for Prosody and the Application of Phonological Processes]. *Sínteses*; 2023. p. 8. Available from: https://revistas.iel.unicamp.br/index.php/sinteses/article/view/6275 [Last accessed on 2024 July 29] [In Portuguese].