IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 June 2022, accepted 22 September 2022, date of publication 3 October 2022, date of current version 10 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3211295

RESEARCH ARTICLE

Clinically Relevant Sound-Based Features in COVID-19 Identification: Robustness Assessment With a Data-Centric Machine Learning Pipeline

PEDRO MATIAS[1], JOÃO COSTA[1], ANDRÉ V. CARREIRO[1], HUGO GAMBOA[1,5], (Senior Member, IEEE), INÊS SOUSA[1], PEDRO GÓMEZ[2], (Life Member, IEEE), JOANA SOUSA[3], NUNO NEUPARTH[4], PEDRO CARREIRO-MARTINS[4,6], AND FILIPE SOARES[1]

[1]Fraunhofer Portugal AICOS–Porto, 4200-135 Porto, Portugal
[2]NeuSpeLab, CTB, Universidad Politécnica de Madrid, 28223 Madrid, Spain
[3]NOS Inovação, Campo Grande, 1600-404 Lisbon, Portugal
[4]Comprehensive Health Research Center (CHRC), NOVA Medical School, 1169-056 Lisbon, Portugal
[5]Laboratory for Instrumentation, Biomedical Engineering and Radiation Physics (LIBPhys), Faculdade de Ciências e Tecnologia, NOVA University of Lisbon, 2829-516 Caparica, Portugal
[6]Allergy and Clinical Immunology Department, Centro Hospitalar Universitário de Lisboa Central (CHULC) EPE, 1150-199 Lisbon, Portugal

Corresponding author: Pedro Matias (pedro.matias@fraunhofer.pt)

This article is a result of the project ConnectedHealth (n.° 46858) and OSCAR (LISBOA-01-02B7-FEDER-051277 | POCI-01-02B7-FEDER-051277), supported by Competitiveness and Internationalisation Operational Programme (POCI) and Lisbon Regional Operational Programme (LISBOA 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF).

ABSTRACT As long as the COVID-19 pandemic is still active in most countries worldwide, rapid diagnostic continues to be crucial to mitigate the impact of seasonal infection waves. Commercialized rapid antigen self-tests proved they cannot handle the most demanding periods, lacking availability and leading to cost rises. Thus, developing a non-invasive, costless, and more decentralized technology capable of giving people feedback about the COVID-19 infection probability would fill these gaps. This paper explores a sound-based analysis of vocal and respiratory audio data to achieve that objective. This work presents a modular data-centric Machine Learning pipeline for COVID-19 identification from voice and respiratory audio samples. Signals are processed to extract and classify relevant segments that contain informative events, such as coughing or breathing. Temporal, amplitude, spectral, cepstral, and phonetic features are extracted from audio along with available metadata for COVID-19 identification. Audio augmentation and data balancing techniques are used to mitigate class disproportionality. The open-access Coswara and COVID-19 Sounds datasets were used to test the performance of the proposed architecture. Obtained sensitivity scores ranged from 60.00% to 80.00% in Coswara and from 51.43% to 77.14% in COVID-19 Sounds. Although previous works report higher accuracy on COVID-19 detection, this research focused on a data-centric approach by validating the quality of the samples, segmenting the speech events, and exploring interpretable features with physiological meaning. As the pandemic evolves, its lessons must endure, and pipelines such as the proposed one will help prepare new stages where quick and easy disease identification is essential.

INDEX TERMS COVID-19, speech, vocal tract, signal processing, feature extraction, data-centric, machine learning.

I. INTRODUCTION

The SARS-CoV-2 Coronavirus is the agent responsible for the COVID-19 respiratory infection, whose transmission among the worldwide community has reached large scale levels over the last two years, which led us into a global pandemic scenario [1]. Up to February 2022, COVID-19 caused more than 418 million confirmed cases and more than 5.8 million deaths worldwide [2]. Infection fatality rates range from 0.1% to 18.1%, always considering that these values can

The associate editor coordinating the review of this manuscript and approving it for publication was Binit Lukose.

vary across different regions, age subgroups, and healthcare support overload levels [3]. Actually, this last point is the problem that governments tackle with the highest priority, i.e., when the amount of people needing medical assistance overcomes the healthcare system response capacity. Although vaccination has helped to protect part of the world's population and control the spread of the virus, immunization levels tend to decrease over time [4], [5], which may contribute to a periodical rise of the transmission rates. This will require a continuous, fast, and accurate screening of the SARS-CoV-2 in order to prevent the possible virus spreading, apart from a vaccination boost (at least until a more effective treatment is validated and made available). Given that the vaccination process is not as fast as desired [6], screening tools earn relevance as being the main instrument to track and restrain the spread of the disease.

Currently, the most accurate and commonly applied SARS-CoV-2 virus diagnostic tools [7] are: molecular (real-time polymerase chain reaction - RT-PCR) and rapid antigen tests. Nonetheless, most of them are not fast-tracking tests, besides being invasive diagnostic techniques and still costly. These high sensitivity tests also result in higher response times, making them far for an ideal use in triage systems requiring fast screening tools to sort patients by treatment priority or even in locations with limited standard diagnostic tests.

Non-invasive COVID-19 screening tools have already been proposed by some research groups, going from categorical data (e.g., symptomatic, demographic-related information) processing models [8], [9], to speech processing models [10], [11], [12], or even a combination of both.

This paper proposes a module-integrated end-to-end framework for COVID-19 screening using audio recordings and Machine Learning (ML) techniques. The proposed pipeline returns the probability of SARS-CoV-2 infection based on speech audio inputs, symptomatic, and demographic information. The feature engineering stage brings out new physiological relevant features from the audio inputs, allowing us to characterize and distinguish a healthy from an infected vocal tract. Some of these features add an explainable layer to the output, resulting in a greater approximation between the model prediction and the physiological phenomena occurring inside the vocal tract of the patients, additionally contributing to a better understanding of the disease. The results are validated through the analysis of two large-scale databases: Coswara [13] and COVID-19 Sounds [14] datasets, through a pipeline designed to handle most of the found data quality-related issues. That is followed by a robustness assessment of our results and a critical comparison with the recent literature. Thus, this framework poses itself as a fast, non-invasive, and interpretable screening methodology for the disease that, with further clinical validation, can offer many advantages over the standard tools already mentioned.

The following Section II presents a brief description of the current state-of-the-art related to COVID-19 and present efforts to tackle it through audio processing. The methodology used is described in Section IV. The obtained results are presented and discussed in Section V and VI, respectively. In Section VII some final remarks and future work steps are described.

II. RELATED WORK

Since the World Health Organization (WHO) set COVID-19 within a global pandemic context, there have been huge efforts made by several companies and research groups around the world concerning the development of fast and effective diagnostic methods [12], [15], and, most recently, long-term post-disease assessment [16], [17]. In this present review, we will converge into diagnostic-based studies whose approaches adopted audio processing techniques combined with ML or Deep Learning (DL) strategies to provide a fair comparison with the proposed methodology. Most of the papers explored open-access crowdsourced databases, since they provide an easier reach to higher volumes of data and would eliminate the need of acquiring new samples, allowing a fast and forward track to the analysis stage.

Brown et al. [18] introduced a feature-based ML technique to explore how distinguishable cough and breathing audio recordings (collected from COVID-19 Sounds dataset) were between COVID-19 positive and healthy subjects. After feeding a set of handcrafted and deep features to the classifier, several classification tasks were followed. The best score was achieved using cough audio features to distinguish COVID-19 positive and negative subjects presenting self-reported cough symptoms (80%, 72% of precision, recall).

Coppock *et al.* [19] proposed an end-to-end custom Convolutional Neural Network (CNN) to detect COVID-19 using breathing and cough audio samples from a 355-participant crowdsourced dataset (a subset of the COVID-19 Sounds dataset). The audio spectrograms were generated to feed the CNN during training. As in the previous study, many tasks (with different combinations of symptomatic and disease categories) were evaluated. The best score has been achieved at discriminating COVID-19 positive and asthma subjects with self-reported cough (90.9% of Area Under the ROC curve, 77.4% Unweighted Average Recall).

A similar framework is presented by Pahar *et al.* [20], that explored ML- and DL-based techniques to extract relevant information from raw cough audios (retrieved from Coswara and Sarcos datasets) to separate COVID-19 positive and negative subjects. A set of handcrafted features was computed and provided to both classical ML and Deep Neural Network (DNN) models. In the classical ML scenario, the best score was reached for a Multi Layer Perceptron (MLP) classifier (about 87% and 88% of specificity and sensitivity, respectively). Regarding the DL evaluation, the best score has been obtained with a ResNet50 (98%, 93% of Specificity and Sensitivity).

Melek [21] implemented an ML-based system to detect COVID-19 patients based on a single cough sound. The data includes recordings from Virufy and NoCoCoDa datasets, comprising a total of 107 COVID-19 positive and

73 negative participants. After extracting the Mel-frequency cepstral coefficients (MFCCs) from the audios, a few different classical ML estimators were tested. Leave-one-out cross-validation and sequential forward selection optimization strategies returned a K-nearest neighbors' model as the best performing (about 94% of accuracy).

Anupam *et al.* [22] introduced a robust tool for fast COVID-19 screening through the analysis of cough audio signals (from the Coswara database) intending to relieve the overwhelming pressure on most hospitals and healthcare facilities. The authors have relied their strategy on an ML pipeline, starting with signal pre-processing, a (temporal, statistical, and spectral) feature extraction stage, and classification (with implicit optimization). The best performance was achieved with an SVM (Support Vector Machine) model (about 97% and 98% of sensitivity and specificity, respectively).

A slightly different approach has been taken by Meister *et al.* [23] who provided an extensive analysis over the most relevant features to extract from audio samples. They ranked 15 audio features (from temporal, spectral, and tempo-spectral domains) evaluated on two large-scale databases (COVID-19 Sounds and Coswara). This ranking step rested on the assumption that unique patterns repeated across independent datasets were COVID-19 and not dataset-specific. Regarding the COVID-19 positive vs. healthy subjects distinction, a Random Forest model trained solely with spectral audio features and a cepstral feature-based SVM have returned the best scores (around 87% of averaged AUC).

A DL-based strategy has been adopted by Hassan *et al.* [24], that extracted information from cough, breathing, and voice audio samples from 80 participants (60 healthy and 20 COVID-19 patients) recruited in different United Arab Emirates hospitals. A set of temporal, spectral, and cepstral hand-coded audio features was computed across these signals and then fed into an Long short-term memory (LSTM) Neural Network model. The highest performance has been achieved for only the evaluation of the breathing signals (98%, 100% of recall/precision).

Table 1 helps summarize the tasks and best performances achieved by each one of the aforementioned studies. Although the majority of these approaches reports good performance scores, they lack at handling the quality of the audios, as well as the interpretability of the output. These implementations focus more at achieving high metric scores, rather than supporting them on clinically relevant information. Hence, that can be a bottleneck at the time of validating the models or simply scaling them into the real-world.

III. SCIENTIFIC BACKGROUND

A. COVID-19 BIOLOGICAL IMPACT ON RESPIRATORY TRACT

Infections of the respiratory tract [25] can be classified according to the symptoms reported and the anatomic structures involved. Two main groups can be defined: Upper and Lower Respiratory Infections (URI and LRI). URIs do not usually produce severe problems in healthy individuals, as they affect only upper respiratory structures, such as the nasal tract, pharynx, larynx, and not the lower airways or lungs. Common cold, sinusitis, and pharyngitis are some of the conditions that originate in the upper respiratory tract. Regarding LRIs, they trigger inflammation on the lungs or the lower airways (bronchi, bronchioles, alveoli), leading to more severe respiratory issues.

Depending on the extent of the inflammation, the damage to the lungs can lead to a reduction in the area available for oxygen exchange with the blood, which in severe cases leads to the systemic dysfunction of the body.

Bronchitis, bronchiolitis, and pneumonia are included within this range of infections triggered in the lower respiratory tract. Bronchitis and bronchiolitis [26] cause inflammation of the bronchi and the smaller distal airways, accompanied by cough, sputum, and wheezing. In contrast, pneumonia [26] affects the alveoli, and the lung parenchyma gets infected, which may cause cough along with breathing problems. Moreover, flu-like infections can also extend up to the lower tract (in severe cases), although most infections only affect the upper tract [27].

Cough-related symptoms are associated with several respiratory infections [28]. A single cough event is composed of three distinct mechanisms: explosive, intermediate, and voiced [29]. It starts with an inspiration phase, glottal closure, production of high thoracic pressure, and explosive air exhalation towards the mouth [30], [31]. Cough events can be voluntary, when forced by the individual, or involuntary, when induced by cough receptors inside the respiratory airways. As the vocal and respiratory tracts get morphologically modified for different infections, the sound produced by the expelled airflow mechanism will also reveal different patterns. The same applies to breathing and speech-related symptoms that appear to show perturbations on COVID-19 infected patients and whose sound patterns are also affected by the respiratory apparatus alterations. The analysis of such biologically-modified sounds may, thus, gain an important role in these scenarios.

The clinical picture of SARS-CoV-2 infection can range from asymptomatic to acute lung injury [32]. Depending on the severity, patients can be categorized as outpatients (not requiring hospitalization) and inpatients (hospitalized). In fact, outpatients end up triggering an effective antiviral immune response with flu-like symptoms, such as fever, dry cough, fatigue, sore throat, loss of smell, and headaches, recovering without developing acute symptoms [32]. However, a small percentage of infected people are inpatients (that may notice more serious symptoms [2] namely shortness of breath, speech loss, chest pain). Their local inflammation becomes exacerbated, possibly progressing to an acute respiratory distress syndrome (ARDS) [33]. Overall, the course of this infection can vary between asymptomatic (no symptoms), mild (recurrent respiratory and

TABLE 1. General synthesis of the performances obtained in the most recent literature, along with the defined tasks and datasets used.

Authors	Dataset	Dimension (COVID-19 / Total subjects)	Task	Best Results
Brown et al. [18]	COVID-19 Sounds	54 / 86	COVID-19 Positive vs Negative (with reported cough)	80% Precision 72% Recall
Coppock et al. [19]	COVID-19 Sounds	62 / 355	COVID-19 Positive vs Asthma (with reported cough)	91% AUC 77% UAR
Pahar et al. [20]	Coswara and Sarco	92 / 1171	COVID-19 Positive vs Healthy	98% Specificity 93% Sensitivity
Melek et al. [21]	Virufy and NoCoCoDa	107 / 180	COVID-19 Positive vs Negative	94% Accuracy
Anupam et al. [22]	Coswara	160 / 640	COVID-19 Positive vs Healthy	98% Specificity 97% Sensitivity
Meister et al. [23]	Coswara and COVID-19 Sounds	81 / 1155 (Coswara) 111 / 305 (COVID-19 Sounds)	COVID-19 Positive vs Healthy	87% AUC
Hassan et al. [24]	Speech Corpus (private)	20 / 80	COVID-19 Positive vs Healthy	100% Precision 98% Recall

AUC - Area Under the ROC Curve; UAR - Unweighted Average Recall

non-respiratory symptoms), moderate (pneumonia-like symptoms), and severe (acute pneumonia) [34].

B. BIOMARKER EXTRACTION THROUGH AUDIO PROCESSING

The previous section has shown that COVID-19 causes respiratory-related symptoms in most of the patients, such as cough, shortness of breath, throat inflammation, among others [35]. Nonetheless, such symptoms are only analyzed as being present or not, through self-report or clinician reports, besides not being followed by any quantified measurement of their intensity, not even in their manifestation patterns within the individual's respiratory tract. Such measures could help characterize the respiratory tract and draw more detailed information about the pathophysiology of this disease. Fortunately, there are many techniques to monitor the airways, including stethoscope auscultation, radiography, and invasive sampling techniques (e.g., induced sputum, endotracheal aspiration, or bronchoscopy [27]). However, these methods are either invasive, qualitative, or time-consuming.

Some literature shows evidence that vocalized breathing is related to perturbations on the anatomy and the physiology of the vocal tract [36]. Rudraraju *et al.* [37] found strong correlations between cough sound features and FEV1, FVC parameters from spirometry exams. Abeyratne *et al.* [38] aimed to distinguish childhood pneumonia from other respiratory diseases through the analysis of cough audio recordings. Bartl-Pokorny *et al.* [39] discovered evidence of discontinuities in the lung airflow during vowel phonation in COVID-19 positive individuals. Depending on the physical conditions and dimensions of the vocal tract, the airflow characteristics (produced by a cough, breath, or any speaking exercise) may generate different sound patterns that (when properly recorded) any microphone can capture. Resorting to detailed statistical, temporal [40], spectral [41], and cepstral [39] analysis of these audio recordings, relevant biomarkers [42] can be identified as distinctive between different respiratory infections or even indicative of their severity. Such analysis can help reconstruct and evaluate (quickly and wisely) the physiological state of the vocal tract, which is currently only possible through more invasive assessments.

For instance, Figure 1 shows how easily wheezings can be detected in a cough audio recording using a spectrogram representation. Through a quick visual inspection, we notice the bottom-left spectrogram contains a series of well-defined horizontal bands (within [1, 2] kHz range), which suggests the presence of wheezing in the considered cough events. The same does not apply to the bottom-right spectrogram, where those bands cannot be identified in the same spectral characterization.

In fact, speech sounds analysis can overcome the invasiveness, cost, and time consumption drawbacks. At the same time, they provide detailed information about the patients' respiratory physiological condition through phonetic fingerprints. As speech disturbances and respiratory problems (shortness of breath, dry cough) are some of the most common symptoms of the COVID-19 disease, it makes sense to explore audio samples enhancing these speech types so that disease-specific biomarkers can be found. These would support a fast and effective disease screening and other useful clinical observations.

IV. METHODOLOGY

A modular ML pipeline was developed to classify subjects as healthy or COVID-positive, based on extracted features of the audio signals from several produced sounds, such as coughing, breathing, and prolonged vowel utterances. Relevant statistical, temporal, and spectral signal features are extracted for each audio signal and grouped by subject to perform classification on the subject's level as a whole.

Figure 2 shows a schematic of the implemented pipeline and its seven main modules, described below and in the following subsections.

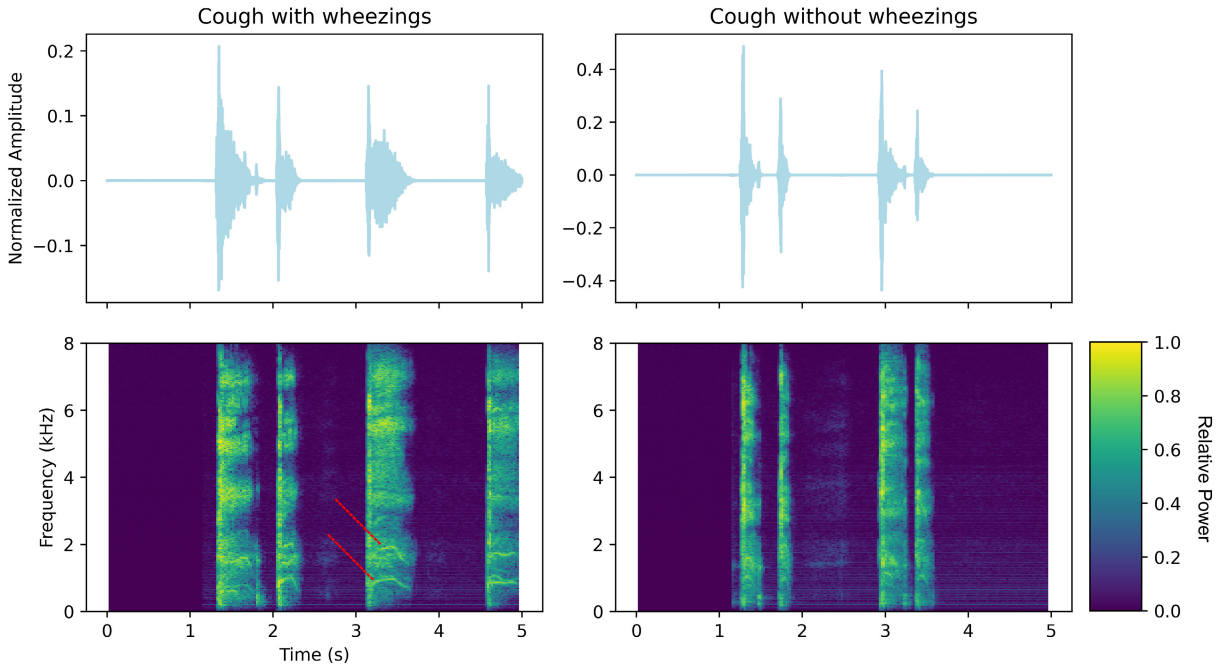


FIGURE 1. Presence and absence of horizontal spectral bands in cough events with and without wheezings, respectively. Top sub-figures represents the raw audio signals, below which the spectrogram is shown.

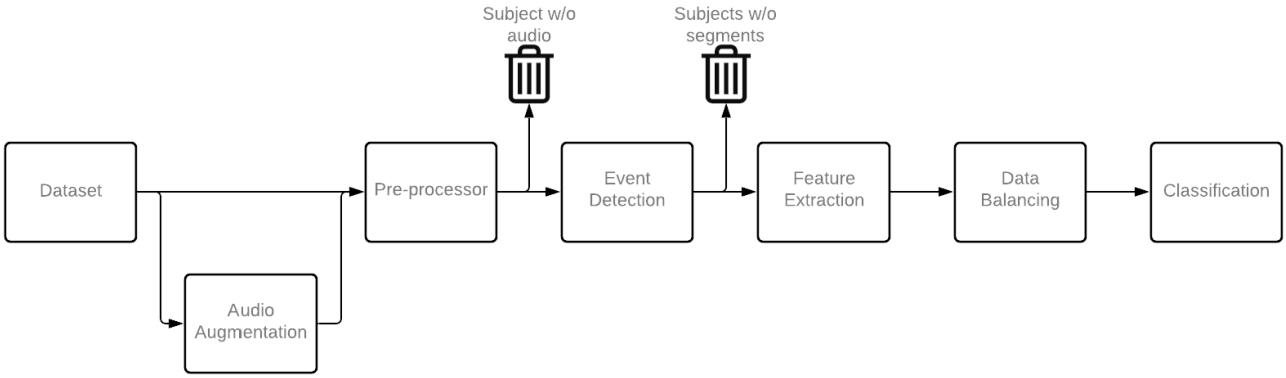


FIGURE 2. Modular pipeline for COVID-19 diagnosis through analysis of speech, cough, and breathing sounds.

The pipeline is composed of seven main modules that are combined to form distinct ML experiments: Dataset, Data Augmentation, Audio Pre-processing, Event Detection, Feature Extraction, Data Balancing, and the Tree-based Pipeline Optimization Tool (TPOT) automated ML pipeline (in turn comprising feature selection, normalization, and classification models).

Note that the performed ML experiments are dataset-dependent, i.e., datasets were not joined for training. The considered datasets resulted from different acquisition protocols, with distinct respiratory and speech sounds recorded, differences in audio quality, number of recorded subjects, and ground truth determination. Thus, the experiments focus on a single dataset each time.

A. DATASETS

1) COSWARA DATASET

The Coswara dataset [13] is a large-scale database of respiratory sounds, namely cough, breath, and voice, released in 2020 and created within the scope of the Coswara project developed by the Indian Institute of Science (IISc) Bangalore. The dataset samples are crowdsourced through a website tool. Nine different speech types are requested during the recordings: cough (shallow, heavy), breath (shallow, deep), sustained vowel phonation (three types; /a:/, /i:/, /u:/), and counting (normal, fast-paced). Alongside the audio inputs, the users provide other symptomatic, demographic, disease-related data, and the covid status whose reliability can be or not supported with a validated test. Considering this dataset is

continuously growing over time, for reproducibility matters, the dataset explored is the version updated on September 30th, 2021, with a total of 2233 unique subjects. Table 2 helps understand the distinct types of labels, our health status definition, and the relative frequency over the complete set of participants.

TABLE 2. Subjects distribution over the set of labels provided in Coswara dataset. Highlighted labels were used to define our health status condition (first column).

Health status	Raw Label	Number of subjects	Relative Frequency
Healthy	**healthy**	**1379**	**0.62**
	recovered_full	99	0.04
	resp_illness_not_identified	151	0.07
	no_resp_illness_exposed	174	0.07
Covid-19 Positive	**positive_moderate**	**118**	**0.05**
	positive_mild	**261**	**0.12**
	positive_asymp	**51**	**0.02**

In ground truth definition terms, only the Coswara's *healthy* label has been assigned to our definition of *Healthy* status. Such a restriction intended to offer higher status reliability for our training/test sets, considering the remaining labels (*recovered_full*, *resp_illness_not_identified*, and *no_resp_illness_exposed*) were not really conclusive at all. On the other hand, all the COVID-19 positive labels (with *moderate*, *mild*, or *no* symptoms) defined our *COVID-19 positive* status.

2) COVID-19 SOUNDS DATASET

The COVID-19 Sounds database is the largest multimodal dataset of respiratory sounds for COVID-19 detection, containing 53,449 audio samples crowd-sourced from 36,116 distinct participants [14]. Multimodal respiratory sources comprise cough, breathing, and voice recordings (sentence reading). Beyond respiratory sounds, additional metadata, such as demographic, symptomatic, and disease-related data, is collected. The COVID-19 status is self-reported, thus not validated by any of the abovementioned tests. The major health-related categories are COVID-19-positive, healthy, and asthmatic (although none are clinically confirmed). After request, a subset of the full dataset was provided for analysis (containing only *breath* and *cough* audios, without any metadata). Table 3 presents the distribution of these different labels over the provided sub-set of participants.

TABLE 3. Subjects distribution over the set of labels provided in COVID-19 Sounds dataset. Highlighted labels were used to define our health status condition (first column).

Health status	Raw Label	Number of subjects	Relative Frequency
Healthy	**healthynosymp**	**298**	**0.61**
	healthywithcough	**32**	**0.06**
	asthmawithcough	20	0.04
Covid-19 Positive	**covidnocough**	**87**	**0.18**
	covidwithcough	**54**	**0.11**

We removed asthmatic participants from our subset because the asthma factor implies a modified respiratory tract and symptoms. Since it could add confusion during the model training stage (given the smaller number of patients), we decided not to mix up our classes' distributions in the sense that these were, pathologically, as distinguishable as possible.

B. SUBJECT REJECTION

Throughout the audio processing pipeline, subjects' audio files undergo different checkpoint steps, confirming their eligibility for the following stage. Such steps are generally the same, with slight differences among both datasets addressed here. The criteria were:

1) **Quality-based manual annotations**: following the efforts made by https://github.com/iiscleap/Coswara-ExpLEAP Lab [13], these annotations allowed us to *a priori* reject some ineligible subjects either due to the presence of any *bad* or *noisy* audios. Such a step only applies to Coswara dataset processing;
2) **Empty files**: subjects containing any empty audio files are immediately set as ineligible for the analysis;
3) **No available segments**: subjects whose audio files don't have any relevant segments or these are wrongly classified by the segmenter are excluded.

As COVID-19 Sounds dataset already undergoes an audio quality check (through an automatic tool) [18], the quality of the provided audio files is assumed to be good enough for analysis.

C. DATA AUGMENTATION

Data augmentation was used to mitigate the unbalanced nature of both datasets regarding the percentage of COVID-19 positive subjects.

Augmented subjects are generated by copying pre-existing subjects altering their audio data. Augmented audio signals are created using the *audiomentations* Python package [43], which transforms the original audio using a combination of techniques, namely time stretching, pitch shifting, time shifting, gain increase, and addition of Gaussian noise.

Note that only the train set can be subjected to data augmentation. Test sets across the several experiments contain no augmented subjects.

D. AUDIO PRE-PROCESSING

After augmentation, all audio signals present in the selected dataset are subject to several pre-processing steps to ease their analysis. The implemented pre-processing steps are similar to the ones used in [44].

To remove high-frequency content, signals are converted to mono, normalized, and low-pass filtered (cutoff frequency of 6 kHz).

E. SPEECH EVENT SEGMENTATION AND CATEGORIZATION

The collected audio samples from different speech types were recruited from crowdsourcing. The total duration of each signal and the interval between the events (e.g., coughs, breathing, counting) were not under control. Moreover, the background environment could differ among participants, so it is not ensured all the samples show an acceptable signal-to-noise ratio. Thus, we decided to perform a segment-wise analysis, extracting and classifying all the relevant (non-silent) segments before undergoing the feature extraction stage. Figure 3 presents a diagram illustrating this speech event categorization pipeline.

Observing Figure 3, we note two consecutive processing stages before feature computation takes place. In Stage 1, the non-silent segments are extracted through a silence trimming tool (from *librosa* python package [45]). In this way, audio silences are removed from our analysis since background sounds can induce undesirable bias both to our feature extraction and model training stages. After extracting the relevant segments, each one is mapped into a spectrogram representation and passed through a pre-trained Speech Event Detection model, which classifies the type of event. Then, segments whose assigned speech event matches the labeled audio sample event are accepted, being rejected otherwise. The window parameters' settings used to extract non-silent segments and generate the spectrograms, as well as other pre-processing restrictions introduced to normalize the input shape, are reported in Table 4.

TABLE 4. Hyperparameters definition for pre-processing audio files regarding non-silent segments extraction and spectrograms generation tasks.

Parameter	Segment extraction	Spectrograms
Window length (s)	0.060	0.040
Window overlap (%)	50	50
Maximum duration (s)	1.0	N.A.
Minimum duration (s)	0.1	N.A.

Summing up, this audio event-based segmentation step helps reject silent (energy-negligible) and non-conformant (not matching the expected audio file label) audio segments.

Relative to the segment classification model selection, our choice fell on a 2D-CNN, after some preliminary experiments with VAE (Variational AutoEncoder), CNN and LSTM architectures. Convolution operations can exploit spatial and temporal correlations of data samples, which explains their extensive use in image- and speech-related applications. In our case, the model was built to distinguish between audio spectrograms generated from different speech types. The designed architecture is represented in Figure 4.

Concerning the model training optimization, hyperparameter tuning was carried out empirically across several experiments, and the best-performing model was selected. The training is dataset-dependent, i.e., there is one single model *per* dataset, meaning that, in this case, a model trained in one dataset was not reused to infer on the other. The selected set of hyperparameters is reported in Table 5 for both datasets.

TABLE 5. Set of hyperparameters applied during the 2D-CNN training step.

Parameter	Value
Epochs	100
Batch size	16
Early Stopping patience	10
Learning-rate	5×10^{-5}
Activation function	SeLU
Convolutional kernel size	3×3

At the inference step, non-silent segments are extracted from the complete audio signal, converted to a spectrogram representation, and passed through the CNN, whose output is then evaluated in a final if-based rule (accept/reject).

F. FEATURE EXTRACTION

Each subject can be described by a set of extracted features obtained from different data sources, such as collected metadata and audio signals from recorded samples. These features form a single feature vector per subject so that classification is performed on a subject level.

1) METADATA FEATURES

A set of features is generated from the available metadata relative to each subject in a dataset. The specific data may vary from dataset to dataset, but feature encoding is fixed.

Categorical variables, such as gender and location, are encoded using a one-hot encoding schema. Yes or no answers, such as the expression or absence of a given symptom, are converted to boolean features. Numerical answers, such as age, number of days since a given test, etc., are used as-is, but when represented as a list of interval categories (e.g., age interval 30-40), the mean value of the interval is used.

2) AUDIO FEATURES

To perform feature extraction, all audio segments are windowed according to customizable window length and overlap parameters since feature computation can be expensive for original audio signals while capturing the collected samples' non-stationary behaviour and evolution. These window parameters can be adjusted according to the type of audio to be analysed, but remain fixed within the same experiment. For instance, it is recommended to use shorter windows for cough signals since they are naturally unstable with high-frequency content.

Audio feature extraction was based on time-series analysis, with several groups of spectral and temporal features calculated for each extracted audio segment window. The *TSFEL* Python package [47] was used as a basis for audio feature extraction, with several relevant features added for this specific application of speech and breathing sound analysis

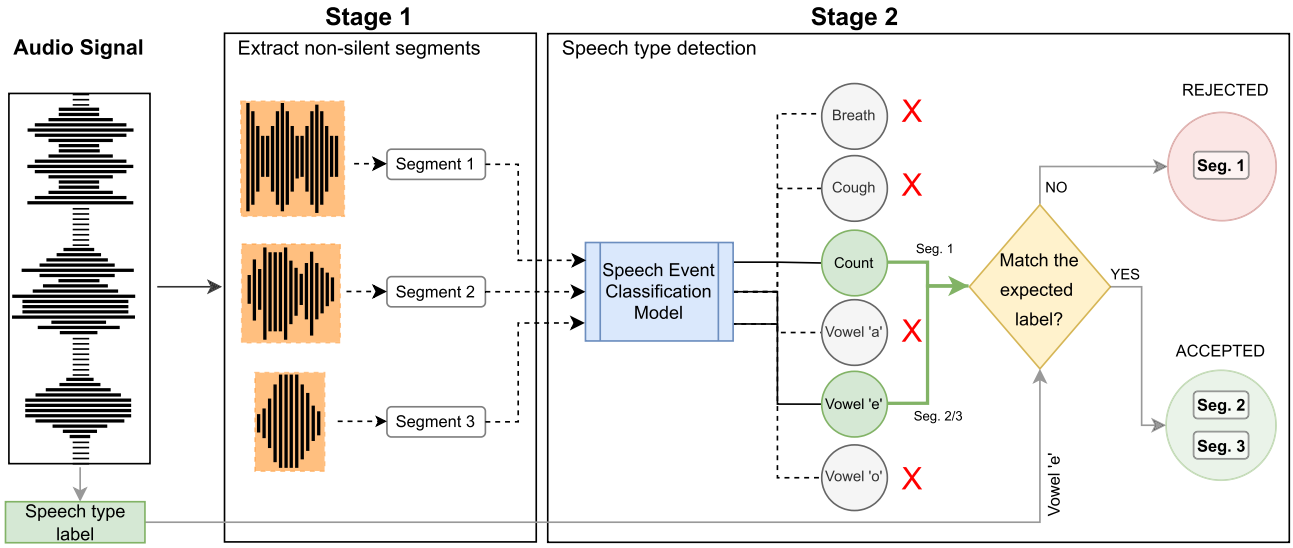


FIGURE 3. Scheme illustrating the speech event detection inference pipeline. Stage 1 comprises the extraction of non-silent segments from the complete audio signal. Stage 2 covers the classification of the audio segments extracted in the previous step, followed by a label matching procedure.

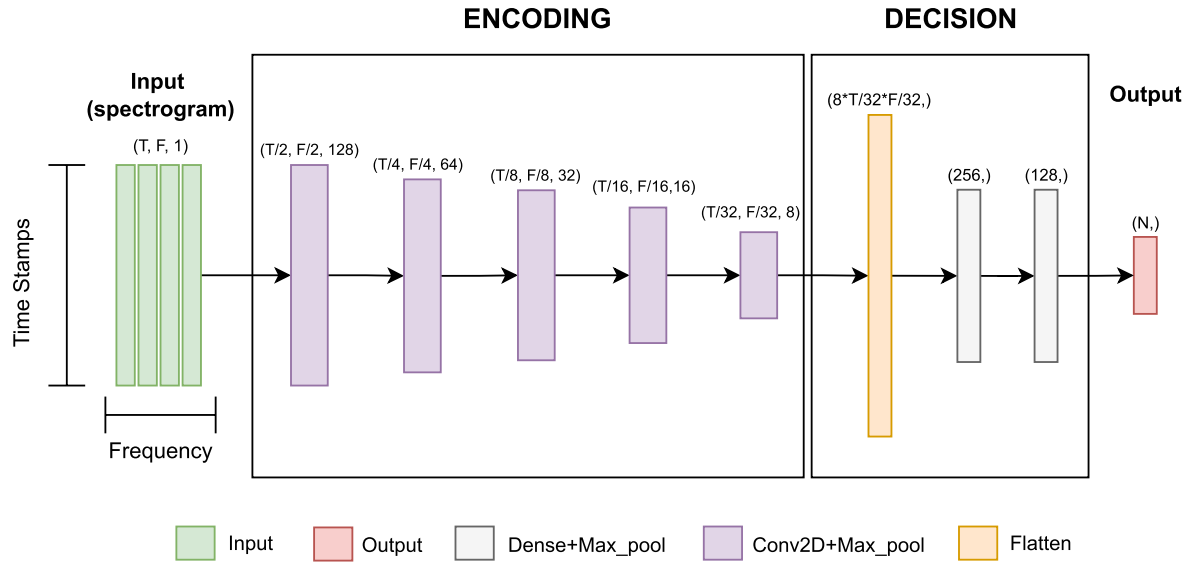


FIGURE 4. Speech event-based classification model. The architecture comprises an encoding part, where 2D-convolutions are computed on the input through five similar consecutive layers, and a decision part, where the convolutional product is used to calculate the confidence level of each defined class. All the neural network layers are followed by a SeLU activation gate [46], except the last one (which has a Softmax probability mapping).

(e.g., phonetic-related features). Table 6 shows a list of features for audio analysis.

3) FEATURE AGGREGATION

The final feature vector for a given subject is given by concatenating the encoded metadata features with the overall audio feature vector. A schema of this operation is shown on Figure 5.

Audio features are extracted on a window level for all segments of all audio types. The set of all window features extracted from all segments categorized as a given audio type is used to describe the features of the referred type, through six descriptive statistics (minimum, maximum, mean, variance, skewness, and kurtosis). Thus, for a subject with K audio types and L_k features extracted from signal windows of audio type k, $\sum_k K \times L_k \times 6$ features compose the overall audio feature vector.

G. DATA BALANCING

The calculated feature matrix may still present a significant unbalance between healthy and COVID-19 subjects even with data augmentation. To combat this phenomenon on a

TABLE 6. List of temporal, spectral and phonetic features used for audio feature extraction. Features marked with * were considered clinically relevant by an expert in pathological speech analysis [48].

Feature type	Number of Features	Feature Names		Reference
Spectral	10	Spectral Centroid Spectral Kurtosis Spectral Spread Spectral Variation Spectral Bandwidth*	Spectral Decrease Spectral Skewness Spectral Slope Spectral Roll-off* Spectral Roll-on	[47]
	2	Period*	Root Mean Square Energy*	[18]
	3	Mean Square Energy Spectral Contrast	Polynomial fit to the Spectrum	[13]
	1	Harmonic to Noise Ratio*		[49]
Cepstral	2	Mel Frequency Cepstral Coefficients*[a]	Linear Prediction Cepstral Coefficients[a]	[47]
	1	Cepstral Peak Prominence*		[50]
Temporal	1	Zero Crossing Rate*		[47]
	1	Jitter*		[51]
Amplitude	1	Shimmer*		[51]
Phonetic	4	Vowel Space Area[b] Bandwidth of Formants*[b]	Formant Centralization Ratio[b] Formants*[b]	[52]

a - first 12 coefficients used
b - first 2 formants used

classifier level, this module allows for data balancing of the feature matrix, where several balancing methods can be used and tested, such as random over/under-sampling and SMOTE [53].

Note that only the train set can be subjected to data balancing techniques. Test sets on all ML experiments are kept with the same class imbalance.

H. TRAINING AND OPTIMIZATION SCHEME

1) DATASET SPLITTING

The dataset has been divided into two subsets: one to train and validate the segment-classification neural network (Subset A), and the other to train, validate, and test the COVID-19 detection model (Subset B). To save as many COVID-19 infected patients as possible (mitigating the discrepancy among the number of healthy vs. COVID participants), we forced Subset A to contain only healthy subjects. Table 7 helps characterizing the content of each subset.

Additionally, Subset B is further randomly split into two distinct ones: B_{train} (used to fit *TPOT* with a cross-validation training scheme) and B_{test} (used to test the *TPOT*'s output model behavior, as described below). The splitting rate was defined as 75% for training and validation, and 25% for testing.

2) CLASSIFICATION OPTIMIZATION

After feature extraction and dataset splitting, data is ready to be fed to a classifier to generate a binary model able to distinguish between healthy and COVID-19 positive subjects based on information extracted from audio files from cough, breathing, and additional metadata.

TABLE 7. Distribution of Coswara dataset subjects over each separated subset.

Subset	Application	% Healthy	% COVID-19
A	Train and validate the speech event detector model	40	0
B	Train, validate, and test the COVID-19 detection model	60	100

The proposed pipeline can be used with any ML classification model as its final module to generate predictions on the COVID-19 status of an unknown subject. For this particular work, the Python package TPOT [54] was used to automatically generate classification pipelines. The training strategy adopted inside TPOT is schematized in Figure 7. Considering our validation set was imbalanced, and the pipeline choice decision would rely on it, the F1-score was defined as the score to maximize (for selecting the best returned pipeline), since it is a fair metric dealing with imbalanced datasets, and it balances well both Sensitivity and Specificity scores from both classes, which is exactly what we are looking for in this task.

The use of TPOT [55], based on genetic programming, allows for quick prototyping of a classification schema, being able to effectively cycle through hundreds of combinations of all available methods in *scikit-learn* for feature selection,

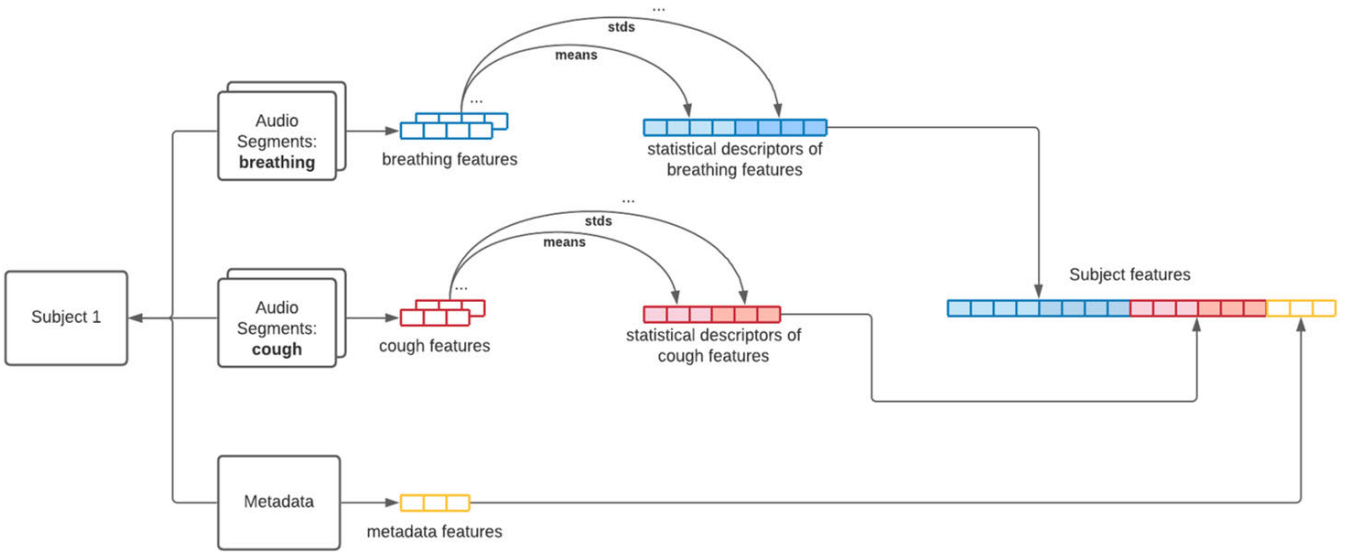


FIGURE 5. Feature extraction and aggregation for a given subject. Audio features are extracted by segments and are summarized with statistical descriptors. Metadata features are obtained by appropriate encoding techniques according to their nature. A final feature vector is obtained by concatenating all audio and metadata features. Note that this example only includes two types of audio files (breathing and coughing), but additional types can be used (for instance, vowels, counting), and contribute to the final subject feature vector.

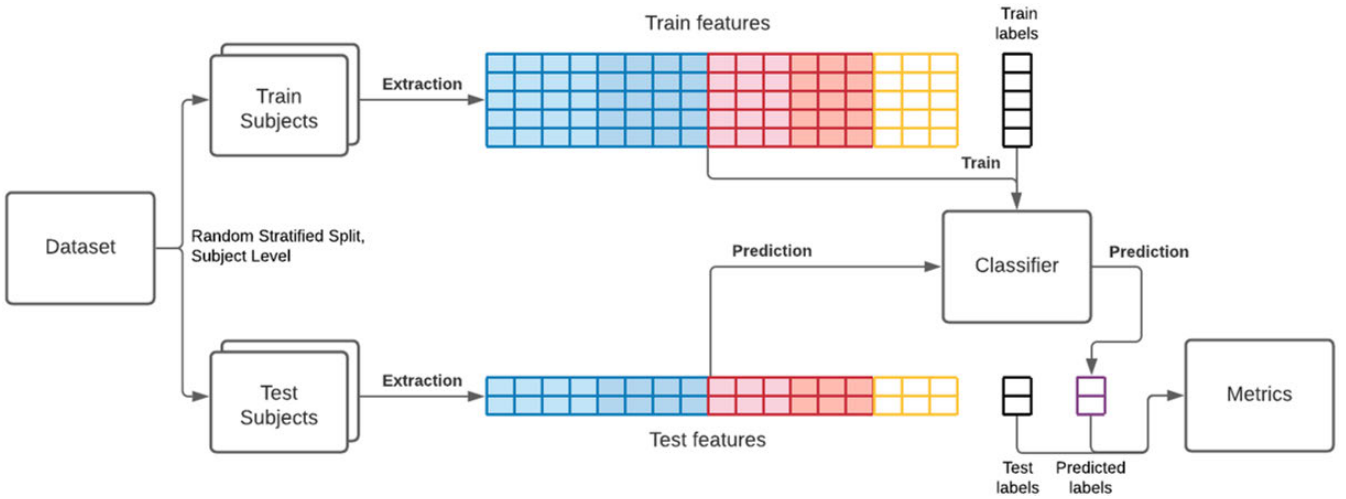


FIGURE 6. Classifier training and performance estimation with the extracted feature matrices, from train and test subsets. The classifier module can be any traditional ML model or an instance for pipeline optimization.

feature transformation, and classification models, while performing cross-validation for hyperparameter tuning of the deployed models for these tasks. Although computationally expensive, this package enables efficient use of the time available to improve the audio processing algorithms. It automatically performs the laborious tasks of feature analysis, removal, and processing, model benchmarking, and hyperparameter tuning. This framework has two critical parameters, that will define the extent of the pipelines' search: *population_size* and *generations*. The number of pipeline configurations being evaluated during TPOT training is given by *population_size* × *generations*. If we take into account a cross-validation training scheme, this number will grow up to *population_size* × *generations* × *n_folds*. TPOT tries a pipeline, evaluates the performance and starts changing parts of that pipeline looking for better performing algorithms. The higher these parameters, the deeper it will search and the longer the running process will take.

After optimization, TPOT returns the best performing pipeline (according to a user-defined metric), with optimized hyperparameters for each step. A Selector-Transformer-Classifier pipeline (comprised of a feature selector, scaler, and a classification model) was used for the ML experiments of this work, with a macro-averaged F1-score as the optimizing metric in a 5-fold cross-validation scheme.

V. RESULTS

This section reports the results obtained from the proposed pipeline regarding training and testing steps applied to the Coswara and COVID-19 Sounds audio samples.

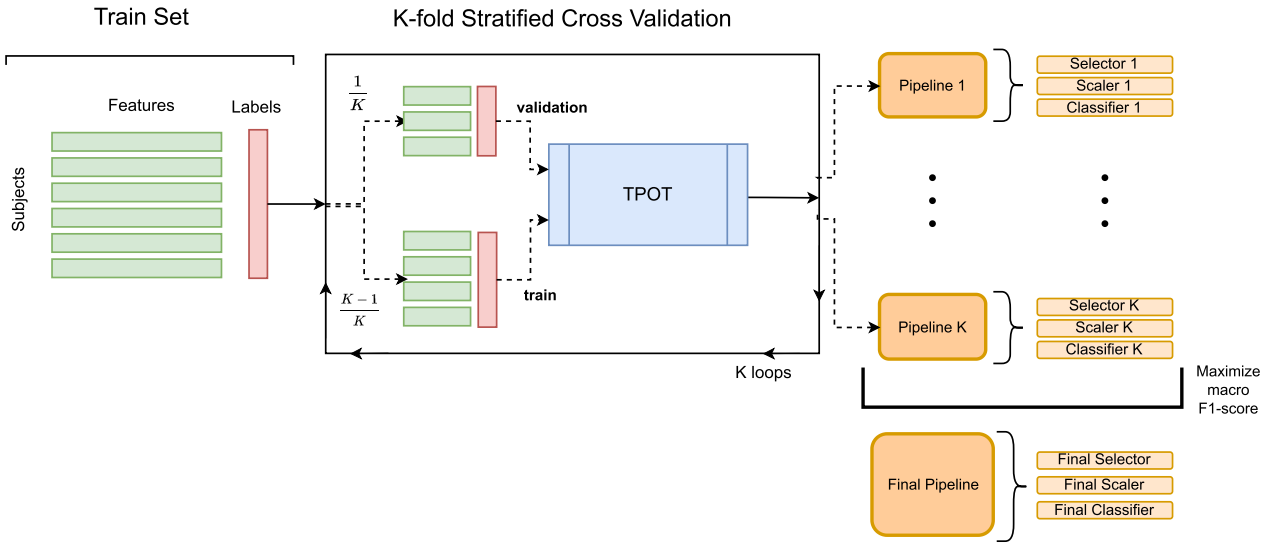


FIGURE 7. Training strategy adopted insite TPOT framework. A K-fold stratified cross-validation flow was combined with a score-based pipeline choice. Each loop returns a pipeline composed by a feature selector, scaler and a classification model. Then, the best pipeline is chosen as the one maximizing the obtained macro F1-score.

A. COSWARA DATASET

1) SPEECH EVENT DETECTOR PERFORMANCE

After pre-processing the audios of each different speech type and subsequently extracting their non-silent segments, these were fed into the speech classification model (CNN), for training. The confusion matrix is presented in Figure 8, and some associated scores are indicated in Table 8. There were a total of 42,719 validation segments.

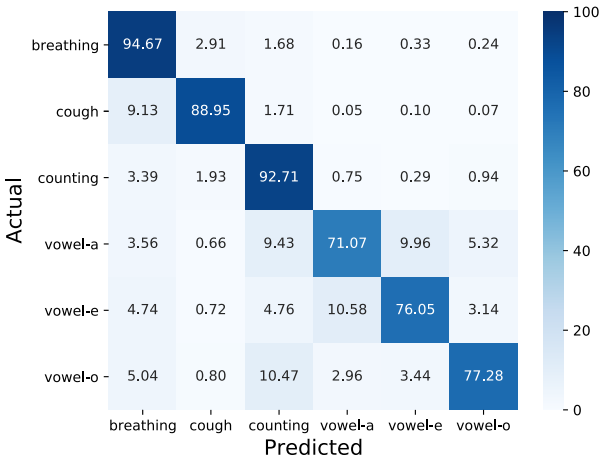


FIGURE 8. Normalized confusion matrix reporting the performance of the speech event classification model in the Coswara dataset.

2) COVID-19 DETECTION MODEL: TRAIN AND TEST

After training the segment classification model, the target segments (from Subset B) are ready to be extracted and go through feature extraction, which will attempt to characterize each subject's health status according to the audios and metadata provided by each subject at the collection phase.

TABLE 8. Performance scores of the speech event detection model applied to the Coswara dataset. Macro scores were chosen instead of any other metric as they are not affected by class balancing issues.

Metric	Score (%)
Macro-Precision	85.70
Macro-Recall	83.46
Macro-F1	84.56

As the subjects are continuously going under conformity check procedures, the number of subjects available at this stage may be lower than the initial number retrieved from the B_{train} subset (see subsection IV-H1). Furthermore, N augmented subjects are generated for each training subject (unless $N = 0$), which helps increase the volume of data to train and fine-tune the *TPOT* chosen model. The general training hyperparameters are defined in Table 10, noting the *Optimization* column refers to the TPOT training scheme (see Figure 7).

Following the subsequent stages, the total number of computed features reached 1,626. In addition, the number of subjects comprehending the test feature sets (on each dataset) is given in Table 11. The models' overall performance over the whole sort of test sets is shown in Table 12. Positive and negative classes are assigned with the *COVID-19 positive* and *Healthy* labels, correspondingly. The best split was obtained when *TPOT* returned a Variance Threshold feature selector, associated with a Standard Scaler normalizer and a Random Forest classifier. The respective confusion matrix is displayed in Figure 10. The complete set of pipelines generated in each iteration during TPOT training stage is introduced in Table 13.

B. COVID-19 SOUNDS DATASET

1) SPEECH EVENT DETECTOR PERFORMANCE

Training the classification model is preceded by audio normalization and segment extraction steps. In this dataset, only *breathing* and *cough* speech types are available, so this problem reduces to a binary classification task. The validation confusion matrix and some associated performance scores, obtained after training, are shown in Figure 9 and Table 9, respectively. The total number of validation segments reached the amount of 3,250.

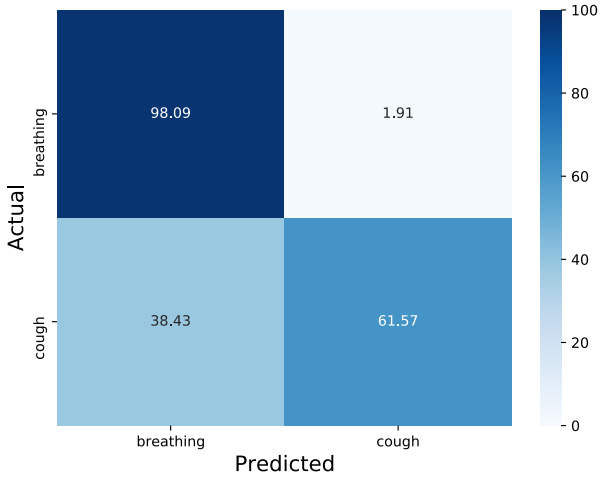


FIGURE 9. Normalized confusion matrix reporting the performance of the speech event classification model in the COVID-19 Sounds dataset.

TABLE 9. Performance scores of the speech event detection model applied to the COVID-19 Sounds dataset. Macro scores were chosen instead of any other metric as they are not affected by class balancing issues.

Metric	Score (%)
Macro-Precision	89.81
Macro-Recall	79.83
Macro-F1	84.53

Due to the resulting performance and some other concerns (addressed in the Section VI), the approach of analysing the complete recordings as one segment (no segment extraction) has been carried out in opposition to what has been done in the Coswara dataset evaluation.

2) COVID-19 DETECTION MODEL: TRAIN AND TEST

The trained speech-event segmenter has not been used in this stage, so Subsets A and B (see Table 7) were reduced to a single set while keeping the same 4:1 train/test splitting ratio. Due to the same reasons reported previously (subjects' rejection criteria), the total number of subjects contained in this Subset might not be the same as the one presented in the dataset description (see Table 3).

Thus, following a similar reasoning path as the one described for Coswara dataset evaluation, the general hyperparameters that returned the best results on the COVID-19 Sounds dataset evaluation are introduced in Table 10. Relative to the feature extraction, the number of features reached 319. The composition of the test set (used to report the classifier performance), concerning the number of gathered subjects, is shown in Table 11. Finally, the overall COVID-19 detection performances are reported in Table 12. The best pipeline generated by *TPOT* was comprised of a Percentile feature selector, a Standard Scaler normalizer, and an MLP classifier. The respective confusion matrix is displayed in Figure 11. The complete set of pipelines generated in each iteration during TPOT training stage is introduced in Table 13.

VI. DISCUSSION

In this section, a critical discussion about the results obtained (either for the speech event and COVID-19 detection models) for both respiratory datasets under analysis is presented. During the interpretation of our classification results, other studies are added into the discussion not to achieve a direct performance comparison (as it was not possible) but to assess our framework's robustness (in terms of data processing and model training good practices) against theirs. Finally, the core elements that also limits the validity of out study are reported in a separate subsection.

A. RESULTS INTERPRETATION

The Coswara dataset has a large volume of samples, though its audio content is sometimes of questionable quality. The Coswara project team led some efforts to mitigate that issue, but because the database keeps growing, that approach becomes unfeasible in the long run. Concerning the present approach, such a high volume of data helped achieve a better convergence during CNN training, softening the influence that sporadic wrong audio segments have in the neural network learning step. In fact, observing the confusion matrix represented in Figure 8, we notice a satisfactory performance, especially considering *breathing*, *cough*, and *counting* classes. Such an overview presented relatively good macro-aggregated scores (see Table 8) accounting for the number of classes, with the F1-score reaching 84.56%. The remaining three classes are vowels, which only differ on the type of vowel and not the speech type. Given that, we consider that the model confusion between vowels is normal and expected, with less impact than other classes (despite occurring as well). Because, in terms of the ground truth definition, the complete audio file label is assigned to all the extracted non-silent segments, this can worsen the validation scores. We cannot ensure that each one of these segments effectively corresponds to the expected or any other speech type (induced by noise, background corruption, or simply human errors).

Hence, the segment-wise approach has been followed considering the performance of the CNN model adequate at distinguishing the segments, offering us an additional degree of certainty that the speech type over which we are extracting features is indeed what we expect. The subsequent feature set was then estimated with *TPOT* and the best-suited pipeline

TABLE 10. General parameters used on the COVID-19 detection model training and optimization scheme for both datasets (Coswara and COVID-19 Sounds) evaluation.

Dataset	Augmentation		Window dynamics for feature extraction [length (ms), overlap (%)]				TPOT settings				Class Balancing
	No. augmented/subject	Class	Breath	Cough	Count	Vowels	Scoring	No. CV splits	Generations	Population size	
Coswara	1	'covid'	[32, 50]	[16, 75]	[32, 50]	[64, 50]	F1-score	5	1	100	OverSampling
COVID-19 Sounds	4										N.A

CV — Cross-Validation

TABLE 11. Composition of the testing feature sets of both Coswara and COVID-19 Sounds datasets in terms of the number of subjects. X_{test} was generate through a random splitting process from the global set of features (X). Mean (μ) and standard deviation (σ) values were calculated over 5 distinct iterations of the same pipeline, regarding each dataset.

Dataset	Number of subjects (X_{test}) $\mu \pm \sigma$	
	COVID-19 positive	Total
Coswara	55.5 $\pm$0.5	195.8 $\pm$1.5
COVID-19 Sounds	35.0 $\pm$0.0	118.0 $\pm$0.0

has been selected. Exceptionally, in this case, computational restrictions affected our iteration runs. As the Coswara dataset comprehends a large number of subjects (each with several data sources to load and process), the complete pipeline execution (with several iterations in a row) became computationally infeasible (due to memory issues), so we opted to pick up several distinct iterations (from different pipeline runs), collect all the metrics, and report the results as for the COVID-19 Sounds dataset. Moreover, for reporting good practices, we decided not to aggregate the results with a mean (μ) and standard deviation (σ) values, as, for each iteration, *TPOT* returns a different pipeline and classifier. Instead, we declared minimum and maximum intervals to evaluate the consistency and variability of the results, apart from the performance. In terms of evaluation, we defined the *COVID-19* recall (sensitivity) and the *Healthy* class precision (NPV) as most relevant under the clinical point of view. In fact, we emphasise a correct classification of all the effective *COVID-19 positive* subjects (high sensitivity) even if, with that, we lose some accuracy on the *Healthy* subjects detection (specificity). If the system is not sensitive enough, we risk labeling a real COVID-19 infected person as *Healthy* and potentiate the spread of the virus.

By analysing Table 12 (top row), we can state not only that the best results obtained are satisfactory but also realistic given the noise variability and non-conformities present in many audio samples. The results also suggest a relative variability either looking at the proposed metrics or the pipeline selected by *TPOT*, which may indicate the performance is weakly dependent on the partition used for testing. Actually, the fact that each subject collected audio samples in uncontrolled environments introduces a new background noise bias that can impair the way the models interpret the data and their subsequent predictions. Therefore, the higher the number of audio sources collected for training the models, the most likely it is to find noisy audios (at least in one source). Moreover, the existing 51 asymptomatic participants that reported positive COVID-19 diagnosis (see Table 2) may be more challenging to detect since they do not present typical symptoms that would effectively reflect their upper/lower respiratory tract condition and the subsequent patterns observed in the recorded audios. Thus, our performance could slightly improve if such a category was removed from our *COVID-19 positive* status.

Concerning the existing literature, we did not find a clear, straightforward comparison as the Coswara dataset is continuously growing sample-wise. Nevertheless, we found recent publications following similar strategies and tasks to those proposed here. Meister et al. [23] applied a 5-fold CV strategy and achieved average COVID-19 detection PPV and Sensitivity scores of 76.70% and 53.09%, respectively, despite having used only *Breath* + *Cough* audios, whereas we included also *voice (vowels, counting)* features to train our models. Nonetheless, our PPV score is slightly lower ([65.00, 73.21] % interval). In contrast, the Sensitivity score reached substantially higher values ([60.00, 80.00] % interval), which means that our model can be more accurate at detecting real COVID-19 infected subjects despite having a marginally higher number of false positives (FP). Such FP can have a greater influence on the PPV score due to the imbalance between classes (there are notably more Healthy than COVID-19 subjects). That can be confirmed through the high NPV score values ([85.14; 91.73] % interval), meaning that the model is rather sensitive to the Healthy class. In any case, our (5-fold CV + test set) evaluation strategy is still more reliable to report on model behaviour, so we believe that our performance is closer to that of the real-world's. Pahar et al. [20] adopted a similar approach by extracting features from *cough* audio signals and additional metadata, feeding them to several ML classifiers, and testing their performance. Although they reported significantly higher scores than ours (reaching 96.73% sensitivity, 97.56% specificity, and 99.16% precision), the authors did not compute the evaluation on a separate test set. As far as we noticed, their performance is reported based on a 10-fold CV strategy, where we still don't know whether the final score was obtained by averaging all the splits or choosing a single one. Despite their good performance (using a reduced dataset), we defend that our algorithm has a more reasonable and practical performance (with a transparent and fair evaluation), given the non-conformities and quality-related issues inherent to this dataset. Anupam et al. [22] have introduced the extraction of a set of statistical, temporal, spectral, and cepstral features

TABLE 12. Classification performance obtained on both datasets (Coswara and COVID-19 Sounds) evaluation. Scores (Sensitivity, Specificity, PPV - positive predictive value, NPV - negative predictive value, Accuracy) are delimited by the minimum and maximum obtained throughout five iterations with different test splits. Positive class - Covid-19 positive; Negative class - Healthy.

Dataset	Results ([Min, Max])				
	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy* (%)
Coswara	[60.00; 80.00]	[84.51; 90.00]	[65.00; 73.21]	[85.14; 91.73]	[75.00; 83.89]
COVID-19 Sounds	[51.43; 77.14]	[83.13; 92.77]	[56.25; 80.00]	[80.23; 89.61]	[67.28; 80.67]

*Accuracy is reported in its balanced format [57]

from cough signals, followed by a training strategy similar to the one described in this paper, with the same 4:1 train/test splitting ratio, also including a data balancing step and a CV optimization scheme. The authors tested a wide range of classifiers, from classical ML to DNN models. We point out this publication due to the massive performance increase when neural networks assume the COVID-19 detection task (from $\sim$ 70.0%/79.1% to $\sim$ 98.0%/91.2% specificity/sensitivity scores). However, even though we consider the results are promising, and the model's validation strategy is fair, not removing the audio silences between cough events introduces the aforementioned background bias that can influence the learning step. We find it challenging handling the signal background in this dataset because the model might be obtaining some information from the environment or the cough rhythm, not having any correlation with the respiratory tract condition, making the results less related to biological causes.

The second experiment described in this paper has been conducted on the COVID-19 Sounds dataset. As the audio samples of this dataset have undergone a quality check procedure, we may expect a larger a priori reliability about their results. Nevertheless, we decided to try our segment-wise approach. In this case, the volume of samples to train the CNN classifier was not as high as in the Coswara dataset.

The multi-class segment classification task has been reduced to a two-class problem. Therefore, a good model convergence and a similar performance might be expected, which was not observed. Analysing Figure 9, the performance scores were much worse than we expected for a binary classification. Macro scores from Table 9 masked such a non-desirable performance, despite being generally worse than those presented in Table 8, regarding the Coswara dataset evaluation. In fact, most of the validation segments were being classified as *breathing* (including a lot of *cough* labeled samples). When listening to the recordings we noted that cough events were frequently preceded by a quick breath (after listening some of the extracted non-silent audio segments). Thus, in our perspective, *cough* recordings contained both *cough* and *breathing* events which, following the described labeling method, caused a corruption of events in the *cough* class more than in the *breathing* class (a *breath* rarely has a *cough* event associated but the opposite is highly plausible). Considering these reasons, in this case, we end up switching to the segment-wise approach in place of an analysis based on the complete recording.

Concerning the COVID-19 detection task, the results from Table 12 (bottom row) have shown that the model presents a better detection of *Healthy* (specificity score ranging from 83.13% to 92.77%) than *COVID-19 positive* subjects (sensitivity score ranging from 51.42% to 77.14%), contrary to what was desired. The PPV and NPV scores confirm this low performance by noticeably presenting smaller (56.25% up to 80.00%) and larger (80.23% up to 89.61%) interval ranges, respectively, indicating that the number of false positives is increased relative to the number of false negatives. However, both classes are substantially imbalanced ($\sim$ 30% *COVID-19 positive* test subjects) forcing the reported tendency (see Table 11). Because the silent regions from the audios were included, some errors during the learning step could result, as these signal components are variable and non-disease related. Also, we believe that a larger sample size would allow for better performances. It would help mitigate the diversity introduced by many external variables (e.g., different recording equipment, background environments). Hence, we consider that, in the case that it had been performed, the segment-wise approach could have helped remove some audio noise sources, improving current results.

In relation to the recent literature, except for one study, the remaining publications were not directly comparable, as the fold/splitting ratio is not provided or the number of overall subjects is different. Brown *et al.* [18] present a COVID-19 detection experiment using exactly the same volume of (*breath* and *cough* audio) data reported in Table 3, a 5:1 train/test splitting (we used 4:1), and the same model validation scheme, using 10 different random seeds (we used 5), each associated with a 5-fold CV + test set evaluation. Their sensitivity scores ranged (approximately) from 52.00% to 80.00%, whereas PPV went from 68.10% to 72.80% (using handcrafted, deep embedded features, or a fusion of both). Observing the Table 12 (bottom row), the results are similar ([51.43, 77.14] % sensitivity and [56.25, 80.00] % PPV intervals) which highlights the competitiveness of our method relative to the one proposed by these authors. Meister et al. [23] have also tested their proposal in the COVID-19 Sounds dataset. They attained sensitivity and PPV average scores of 81.39% and 87.61%. Such a performance is remarkably higher than the results of this work. Again, their evaluation, based only on 5-fold CV results, is not the most reliable, given that there was no separate test set to report the performance scores. Furthermore, the number of samples defined in our paper is also different from their study, making the conditions

not comparable. Coppock *et al.* [19] carried out a COVID-19 detection experiment using DNNs. The authors worked with the same volume of data and a similar validation scheme (optimization on a validation set and performance reporting on a test set) with 3-folds of distinct train, validation, and test sets. The final model prediction UAR (unweighted averaged recall) scores ranged approximately from 72.70% to 82.30% (from their DL-based proposal) and from 62.80% to 73.6% (from a classic ML baseline). Therefore, to keep the same metrics to compare these performances, our proposal's UAR score ranged from 67.28% to 80.67%. It seems that our pipeline slightly outperforms the ML baseline, despite it stands below the DL approach. Nonetheless, both performances seem pretty much similar, yet the presence of a few more common metrics would yield a better comparison amongst these works.

Overall, the proposed framework places itself as bringing a few more advantages than the recent literature studies presented in here. First, it is designed in modules, so that the model training gets independent on any data source. Second, the segment-wise analysis enables the removal of non-relevant audio chunks from the analysis (e.g., silent, noisy), which mitigates the influence that the recording surrounding environment has in the features meaningfulness. Finally, by using a pre-trained speech event classifier allows us to mitigate even more the presence of mislabeled or noisy audio chunk previously assigned to a specific speech type. The combination of these three main elements along with a detailed model training and optimization scheme proves this pipeline follows good and fair data processing and ML practices.

B. LIMITATIONS OF THE STUDY

There are some conditions that restrict our study's validity. One of them is related to the sample size, where a more representative set of data would enable us to achieve a better model convergence onto the real-world scenario, to report more realistic results, and better validate them in a larger testing set. There is also a substantial imbalancing between both (*COVID-19* and *Healthy*) classes which can make the generalization capabilities of the trained model more challenging. The fact both datasets also have different available data sources, the type of computer feature not always match with each other, which ends up hampering the proposed framework's generalization capability. That can be seen as an advantage for training since the proposed modular pipeline is not expecting for a fixed number of data sources, being capable to handle as many sources as there are.

Additionally, the crowd-sourced nature of the data introduces some variability drifts that compromise either the performance scores and the explainability of the output, resulting in a doubtful meaning of the information extracted from the audio data. The fact that the ground truth labels are self-reported and not validated by any certified specialist may also bias the considered health status conditions. Also, adding more symptomatic-related data will enhance the dataset's robustness and better characterize the individuals' health status, supporting the validity of the features extracted from the audio signals.

Respecting the audio data analysis itself, we could have tried any deep learning based approach, as we knew they have shown really promising performances in the recent literature. Nonetheless, we were also aware that it would imply a loss of interpretability, so that option was intentionally not adopted. Moreover, a different limitation may arise concerning the strategy adopted to gather all the speech sound types. While it is extremely useful to gather all the different vocal audio files in the same model training stage (it enriches the model decision capabilities), it also adds a dependency on every single audio file, in the sense any missing/noisy audio will ruin the algorithm. Considering different data fusion techniques must be carried out in a further extension of this work.

VII. CONCLUSION

A. MAJOR ACHIEVEMENTS

In this paper, a new ML-based COVID-19 screening tool has been proposed. The main motivation focused on providing a non-invasive, costless way for early infection detection, whose results could be confirmed later on with a clinically validated test. This would contribute to a more decentralized screening of the disease, which could help release some pressure on the healthcare systems (especially at the most overwhelming times, e.g., fall and winter seasons). Furthermore, additional target points intent to assess and discuss the robustness of this pipeline when evaluated in distinct test sets, as well as to compare the proposed strategy with the recent literature.

To achieve that objective, we relied on different types of speech sounds (*cough*, *breath*, and *voice*) to characterize the subjects' respiratory tract and, together with additional metadata, build a decision-support system capable of computing a final health status report. Nonetheless, the non-desirable quality of some audio samples led us to design a segment-wise analysis, where non-silent segments were extracted and further processed. Furthermore, we also implemented a speech-event detector to increase the degree of certainty about the type of sound present in each of the retrieved segments. It helped assure the meaning and conformity of the information obtained in the feature extraction stage and subsequent decisions.

The feature extraction stage was carefully devised, considering the type of processed audios. The complete feature set comprised a combination of spectral, cepstral, amplitude, temporal, and phonetic indicators reported in the literature, some of them clinically relevant regarding the physiological perspective. In fact, another contribution rested on the *a priori* selection of features, where the introduction of phonetic indicators supported on clinical expertise enabled a more physiological justification of the obtained results and a better characterization of the vocal tract.

The full pipeline was evaluated in two different experiments performed on independent datasets: Coswara and

COVID-19 Sounds. Regarding the Coswara dataset, the speech-event detector has shown its usefulness at validating the content type of each extracted audio segment (84.53% macro F1-score). The COVID-19 detection results were pretty encouraging considering the 55/196 proportion of COVID-19 positive subjects. These results surely support the need to fine-tune the proposed methodology to be validated on additional datasets. Tackling the aforementioned audio quality-related issues would also help to improve the performances and be one step closer to scaling the pipeline onto a realistic framework. Respecting the COVID-19 Sounds dataset, the COVID-19 detection scores got a little worse, suggesting some of the following causes: (1) the small sample size, (2) the lower number of data sources (only *breath* and *cough* audios were made available), (3) the absence of a speech-event detector to assure segment conformity, and the consequent replacement of the proposed segment-wise analysis. We believe that addressing at least one of these concerns would improve scores since the quality of the audios (e.g., presence of background noise) seemed to be much higher in this case.

In conclusion, this first approach has revealed how challenging the analysis of crowdsourcing audio-based datasets can be, especially considering that samples are collected under uncontrolled conditions. The generalization process is complex since non-disease-related variability sources can impact the model learning stage. In terms of evaluation, beyond a single training-test split, this work shows the inherent variability of different data partitions, which may indicate the need for more standardized datasets to validate this type of model. We consider that the proposed pipeline could mitigate some of the reported issues, mainly through the proposed segment-wise analysis (less sensitive to audio background noise), whose results seem promising.

B. FUTURE WORK

A significant source of variability when analysing respiratory and speech sounds from a given population is the uniqueness of each subject's voice that arises from anatomical, environmental, and demographical differences, such as gender, age, habits, and health conditions, among other factors. All these factors impact the intrinsic features of voice, such as pitch, formants, timbre, and may express themselves in the proposed extracted features for COVID-19 diagnosis through audio analysis.

Thus, a logical next step to improve the reported results is to model further the population and the ML algorithms used according to self-reported demographical data, such as gender or age. With these facts in mind, each sub-population used to train a specific model for a particular group will have to deal with less inter-subject variability, thus improving overall results.

Furthermore, to deal with missing metadata, an automatic voice classifier (low-pitch vs. high-pitch) can be used for this purpose, grouping a collected dataset according to intrinsic characteristics of the subject's voice. Several attempts were made, in this work, to implement such a system and improve results based on formant analysis of recorded vowel production. However, due to the crowdsourced nature of the collected datasets, samples were frequently corrupted by background noise (from stationary and non-stationary sources) which made this automatic separation of the dataset way more challenging and inefficient. This is, nonetheless, a relevant field of research to be explored and studied in the future, not only for COVID-19 assisted diagnosis but for audio-based analysis of other respiratory diseases.

Besides that, an interesting extension of this work could comprise an ablation study of the features extracted from the audios, in order to offer a more detailed perspective of each feature's impact on the model decision making process. That could be achieved by introducing different ranking techniques, combine the approaches' results and figure out which features are more recurrently selected as having impact in the model decision. Another trending topic that an extension of this work could fit into is the prediction of the disease severity. As we are currently reaching a more endemic phase of the COVID-19 disease, it is of great interest being capable to discretize better the binary diagnostic space into a severity-driven space. A simple approach could rest on evaluating how the binary classification performance changes by training and/or testing with different levels of disease severity.

Following the limitations of variability and quality in the present study with public datasets (and also in some works in the literature), the authors proposed a new protocol for data collection being currently executed in the Portuguese speaking population in Portugal, with more reliable labelling of positive cases based on certified RT-PCR testing.

APPENDIX A
SUPPLEMENTARY MATERIAL

In this section, a more detailed view over the performance of the COVID-19 detection model is presented either in the form of confusion matrices and pipelines returned by TPOT.

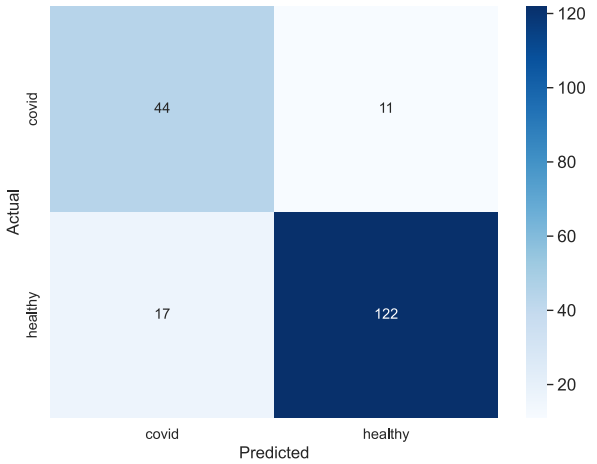


FIGURE 10. Confusion matrix reporting the COVID-19 detection model best obtained performance regarding Coswara dataset evaluation.

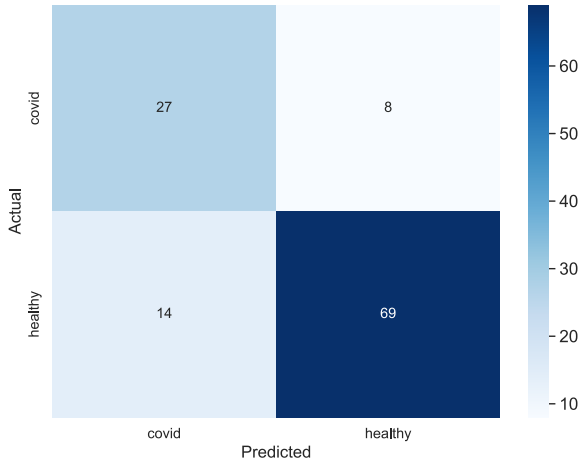


FIGURE 11. Confusion matrix reporting the COVID-19 detection model best obtained performance regarding COVID-19 Sounds dataset evaluation.

TABLE 13. Pipelines returned by TPOT on each iteration performed over each evaluated dataset. Best performing pipelines obtained for each dataset are highlighted.

Dataset	TPOT choice			
	Iter.	Selector	Scaler	Classifier
Coswara	1	RFE (ExtraTrees)	Min Max	Random Forest
	2	Variance Thresh.	Min Max	Gradient Boost
	3	**Variance Thresh.**	**Standard**	**Random Forest**
	4	Variance Thresh.	Min Max	Extra Trees
	5	Select Percentile	Min Max	Random Forest
COVID-19 Sounds	**1**	**Select Percentile**	**Standard**	**MLP**
	2	RFE (Extra Trees)	Standard	MLP
	3	RFE (Extra Trees)	One Hot Encoder	Extra Trees
	4	Select FwE	Max Absolute	Gradient Boost
	5	RFE (Extra Trees)	Max Absolute	Random Forest

FwE — Family-wise Error rate; RFE — Recursive Feature Elimination

ACKNOWLEDGMENT

(Pedro Matias and João Costa contributed equally to this work.)

REFERENCES

[1] World Health Organization. (2021). *Naming the Coronavirus Disease (COVID-19) and the Virus That Causes It*. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it

[2] W. H. Organization. (2021). *Coronavirus*. [Online]. Available: https://www.who.int/health-topics/coronavirus#tab=tab_1

[3] Johns Hopkins University & Medicine. (2022). *Mortality Analyses*. [Online]. Available: https://coronavirus.jhu.edu/data/mortality

[4] N. Andrews, E. Tessier, J. Stowe, C. Gower, F. Kirsebom, R. Simmons, E. Gallagher, M. Chand, K. Brown, S. N. Ladhani, M. Ramsay, and J. L. Bernal, "Vaccine effectiveness and duration of protection of comirnaty, vaxzevria and spikevax against mild and severe COVID-19 in the UK," *medRxiv*, Oct. 2021. [Online]. Available: https://www.medrxiv.org/content/early/2021/10/06/2021.09.15.21263583

[5] P. Nordström, M. Ballin, and A. Nordström, "Effectiveness of COVID-19 vaccination against risk of symptomatic infection, hospitalization, and death up to 9 months: A Swedish total-population cohort study," *Social Sci. Res. Netw.*, vol. 399, no. 10327, pp. 814–823, Oct. 2021, doi: 10.1016/S0140-6736(22)00089-7.

[6] E. Mathieu, H. Ritchie, E. Ortiz-Ospina, M. Roser, J. Hasell, C. Appel, C. Giattino, and L. Rodés-Guirao, "A global database of COVID-19 vaccinations," *Nature Hum. Behaviour*, vol. 5, no. 7, pp. 947–953, Jul. 2021. [Online]. Available: https://www.nature.com/articles/s41562-021-01122-8

[7] D. Trejo Pizzo and S. Esteban, "IATos: AI-powered pre-screening tool for COVID-19 from cough audio samples," 2021, *arXiv:2104.13247*.

[8] Y. Zoabi, S. Deri-Rozov, and N. Shomron, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," *NPJ Digital Medicine*, vol. 4, no. 1, pp. 1–5, Jan. 2021. [Online]. Available: https://www.nature.com/articles/s41746-020-00372-6

[9] W. T. Li, J. Ma, N. Shende, G. Castaneda, J. Chakladar, J. C. Tsai, L. Apostol, C. O. Honda, J. Xu, L. M. Wong, T. Zhang, A. Lee, A. Gnanasekar, T. K. Honda, S. Z. Kuo, M. A. Yu, E. Y. Chang, M. R. Rajasekaran, and W. M. Ongkeko, "Using machine learning of clinical data to diagnose COVID-19: A systematic review and meta-analysis," *BMC Med. Informat. Decis. Making*, vol. 20, no. 1, p. 247, Sep. 2020, doi: 10.1186/s12911-020-01266-z.

[10] G. Deshpande, A. Batliner, and B. W. Schuller, "AI-based human audio processing for COVID-19: A comprehensive overview," *Pattern Recognit.*, vol. 122, Feb. 2022. Art. no. 108289. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320321004696

[11] E. A. Mohammed, M. Keyhani, A. Sanati-Nezhad, S. H. Hejazi, and B. H. Far, "An ensemble learning approach to digital corona virus preliminary screening from cough sounds," *Sci. Rep.*, vol. 11, no. 15404, pp. 1–11, Dec. 2021. [Online]. Available: http://www.nature.com/articles/s41598-021-95042-2

[12] C. Robotti *et al.*, "Machine learning-based voice assessment for the detection of positive and recovered COVID-19 patients," *J. Voice*, Nov. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S089219972100388X

[13] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. R. Chetupalli, P. K. Ghosh, and S. Ganapathy, "Coswara—A database of breathing, cough, and voice sounds for COVID-19 diagnosis," in *Proc. Interspeech*, 2020, pp. 4811–4815, doi: 10.21437/Interspeech.2020-2768.

[14] T. Xia, D. Spathis, C. Brown, J. Ch, A. Grammenos, J. Han, A. Hasthanasombat, E. Bondareva, T. Dang, A. Floto, P. Cicuta, and C. Mascolo, "COVID-19 sounds: A large-scale audio dataset for digital respiratory screening," in *Proc. Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track (Round)*, Aug. 2021, pp. 1–13. [Online]. Available: https://openreview.net/forum?id=9KArJb4r5ZQ

[15] A. Shazia, T. Z. Xuan, J. H. Chuah, J. Usman, P. Qian, and K. W. Lai, "A comparative study of multiple neural network for detection of COVID-19 on chest X-ray," *EURASIP J. Adv. Signal Process.*, vol. 2021, no. 1, Dec. 2021, doi: 10.1186/s13634-021-00755-1.

[16] T. Dang, J. Han, T. Xia, D. Spathis, E. Bondareva, C. Siegele-Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, R. A. Floto, P. Cicuta, and C. Mascolo, "Exploring longitudinal cough, breath, and voice data for COVID-19 progression prediction via sequential deep learning: Model development and validation," *J. Med. Internet Res.*, vol. 24, no. 6, Jun. 2022, Art. no. e37004. [Online]. Available: https://www.jmir.org/2022/6/e37004

[17] W. K. Loo, K. Hasikin, A. Suhaimi, P. L. Yee, K. Teo, K. Xia, P. Qian, Y. Jiang, Y. Zhang, S. Dhanalakshmi, M. M. Azizan, and K. W. Lai, "Systematic review on COVID-19 readmission and risk factors: Future of machine learning in COVID-19 readmission studies," *Frontiers Public Health*, vol. 10, pp. 1–11, May 2022, doi: 10.3389/fpubh.2022.898254.

[18] 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. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. Virtual Event, CA, USA: ACM, Aug. 2020, pp. 3474–3484.

[19] 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 Innov.*, vol. 7, no. 2, pp. 356–362, Apr. 2021. [Online]. Available: https://innovations.bmj.com/lookup/doi/10.1136/bmjinnov-2021-000668

[20] M. Pahar, M. Klopper, R. Warren, and T. Niesler, "COVID-19 cough classification using machine learning and global smartphone recordings," *Comput. Biol. Med.*, vol. 135, Aug. 2021, Art. no. 104572. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521003668

[21] M. Melek, "Diagnosis of COVID-19 and non-COVID-19 patients by classifying only a single cough sound," *Neural Comput. Appl.*, pp. 1–12, Jul. 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8323961/

[22] A. Anupam, N. J. Mohan, S. Sahoo, and S. Chakraborty, "Preliminary diagnosis of COVID-19 based on cough sounds using machine learning algorithms," in *Proc. 5th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, May 2021, pp. 1391–1397.
[23] J. A. Meister, K. An Nguyen, and Z. Luo, "Audio feature ranking for sound-based COVID-19 patient detection," 2021, *arXiv:2104.07128*.
[24] A. Hassan, I. Shahin, and M. B. Alsabek, "COVID-19 detection system using recurrent neural networks," in *Proc. Int. Conf. Commun., Comput., Cybersecurity, Informat. (CCCI)*, Nov. 2020, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/9256562
[25] P. V. Dasaraju and C. Liu, "Infections of the respiratory system," in *Medical Microbiology*, S. Baron, Ed., 4th ed. Galveston, TX, USA: University of Texas Medical Branch at Galveston, 1996. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK8142/
[26] K. Subbarao and S. Mahanty, "Respiratory virus infections: Understanding COVID-19," *Immunity*, vol. 52, no. 6, pp. 905–909, Jun. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1074761320302120
[27] I. I. Bogoch, J. R. Andrews, K. C. Zachary, and E. L. Hohmann, "Diagnosis of influenza from lower respiratory tract sampling after negative upper respiratory tract sampling," *Virulence*, vol. 4, no. 1, pp. 82–84, Jan. 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3544752/
[28] A. Balbani, "Cough: Neurophysiology, methods of research, pharmacological therapy and phonoaudiology," *Int. Arch. Otorhinolaryngol.*, vol. 16, no. 2, pp. 259–268, Apr. 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4435438/
[29] K. K. Lee, S. Matos, K. Ward, G. F. Rafferty, J. Moxham, D. H. Evans, and S. S. Birring, "Sound: A non-invasive measure of cough intensity," *BMJ Open Respiratory Res.*, vol. 4, no. 1, May 2017, Art. no. e000178, doi: 10.1136/bmjresp-2017-000178.
[30] W. Thorpe, M. Kurver, G. King, and C. Salome, "Acoustic analysis of cough," in *Proc. 7th Austral. New Zealand Intell. Inf. Syst. Conf.*, Nov. 2001, pp. 391–394. [Online]. Available: https://ieeexplore.ieee.org/document/974110
[31] J. Korpás, J. Sadlonová, and M. Vrabec, "Analysis of the cough sound: An overview," *Pulmonary Pharmacol.*, vol. 9, nos. 5–6, pp. 261–268, Oct. 1996, doi: 10.1006/pulp.1996.0034.
[32] R. B. Polidoro, R. S. Hagan, R. de Santis Santiago, and N. W. Schmidt, "Overview: Systemic inflammatory response derived from lung injury caused by SARS-CoV-2 infection explains severe outcomes in COVID-19," *Frontiers Immunol.*, vol. 11, Jun. 2020, Art. no. 1626. [Online]. Available: https://www.frontiersin.org/article/10.3389/fimmu.2020.01626
[33] K. Qi, W. Zeng, M. Ye, L. Zheng, C. Song, S. Hu, C. Duan, Y. Wei, J. Peng, W. Zhang, and J. Xu, "Clinical, laboratory, and imaging features of pediatric COVID-19: A systematic review and meta-analysis," *Medicine*, vol. 100, no. 15, 2021, Art. no. e25230. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1477893920300910
[34] J. Mika, J. Tobiasz, J. Zyla, A. Papiez, M. Bach, A. Werner, M. Kozielski, M. Kania, A. Gruca, D. Piotrowski, B. Sobala-Szczygieł, B. Włostowska, P. Foszner, M. Sikora, J. Polanska, and J. Jaroszewicz, "Symptom-based early-stage differentiation between SARS-CoV-2 versus other respiratory tract infections—Upper silesia pilot study," *Sci. Rep.*, vol. 11, no. 1, Jun. 2021. [Online]. Available: https://www.nature.com/articles/s41598-021-93046-6
[35] J. N. Al-Swiahb and M. A. Motiwala, "Upper respiratory tract and otolaryngological manifestations of coronavirus disease 2019 (COVID-19): A systemic review," *SAGE Open Med.*, vol. 9, Jan. 2021, Art. no. 205031212110169, doi: 10.1177/20503121211016965.
[36] J. E. Huber and E. T. Stathopoulos, "Speech breathing across the life span and in disease," in *The Handbook Speech Production*. Hoboken, NJ, USA: Wiley, 2015, pp. 11–33. [Online]. Available: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118584156.ch2 and https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118584156.ch2
[37] G. Rudraraju, S. Palreddy, B. Mamidgi, N. R. Sripada, Y. P. Sai, N. K. Vodnala, and S. P. Haranath, "Cough sound analysis and objective correlation with spirometry and clinical diagnosis," *Informat. Med. Unlocked*, vol. 19, Jan. 2020, Art. no. 100319. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914819304071
[38] U. R. Abeyratne, V. Swarnkar, A. Setyati, and R. Triasih, "Cough sound analysis can rapidly diagnose childhood pneumonia," *Ann. Biomed. Eng.*, vol. 41, no. 11, pp. 2448–2462, Jun. 2013.
[39] K. D. Bartl-Pokorny, F. B. Pokorny, A. Batliner, S. Amiriparian, A. Semertzidou, F. Eyben, E. Kramer, F. Schmidt, R. Schönweiler, M. Wehler, and B. W. Schuller, "The voice of COVID-19: Acoustic correlates of infection in sustained vowels," *J. Acoust. Soc. Amer.*, vol. 149, no. 6, pp. 4377–4383, Jun. 2021.
[40] R. X. A. Pramono, S. A. Imtiaz, and E. Rodriguez-Villegas, "A cough-based algorithm for automatic diagnosis of pertussis," *PLoS ONE*, vol. 11, no. 9, Sep. 2016, Art. no. e0162128, doi: 10.1371/journal.pone.0162128.
[41] K. Anderson, Y. Qiu, A. R. Whittaker, and M. Lucas, "Breath sounds, asthma, and the mobile phone," *Lancet*, vol. 358, no. 9290, pp. 1343–1344, Oct. 2001, doi: 10.1016/S0140-6736(01)06451-0.
[42] E. Maor, N. Tsur, G. Barkai, I. Meister, S. Makmel, E. Friedman, D. Aronovich, D. Mevorach, A. Lerman, E. Zimlichman, and G. Bachar, "Noninvasive vocal biomarker is associated with severe acute respiratory syndrome coronavirus 2 infection," *Mayo Clinic Proceedings: Innov., Quality Outcomes*, vol. 5, no. 3, pp. 654–662, Jun. 2021, doi: 10.1016/j.mayocpiqo.2021.05.007.
[43] I. Jordal. (Feb. 2022). *Audiomentations*. [Online]. Available: https://github.com/iver56/audiomentations
[44] L. Orlandic, T. Teijeiro, and D. Atienza, "The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms," 2020, *arXiv:2009.11644*.
[45] B. McFee, A. Metsai, M. McVicar, S. Balke, C. Thomé, C. Raffel, F. Zalkow, A. Malek, Dana, K. Lee, O. Nieto, D. Ellis, J. Mason, E. Battenberg, and S. Seyfarth. (Feb. 2022). *Librosa/Librosa: 0.9.0*. [Online]. Available: https://zenodo.org/record/5996429
[46] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," 2017, *arXiv:1706.02515*.
[47] M. Barandas, D. Folgado, L. Fernandes, S. Santos, M. Abreu, P. Bota, H. Liu, T. Schultz, and H. Gamboa, "TSFEL: Time series feature extraction library," *SoftwareX*, vol. 11, Jan. 2020, Art. no. 100456. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352711020300017
[48] P. Gómez, J. Mekyska, A. Gómez, D. Palacios, V. Rodellar, and A. Álvarez, "Characterization of Parkinson's disease dysarthria in terms of speech articulation kinematics," *Biomed. Signal Process. Control*, vol. 52, pp. 312–320, Jul. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809419301284
[49] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *IFA Proc.*, vol. 17, no. 1193, pp. 97–110, 1993.
[50] J. Delgado-Hernández, N. M. León-Gómez, L. M. Izquierdo-Arteaga, and Y. Llanos-Fumero, "Cepstral analysis of normal and pathological voice in Spanish adults. Smoothed Cepstral peak prominence in sustained vowels versus connected speech," *Acta Otorrinolaringologica Espanola*, vol. 69, no. 3, pp. 134–140, Jun. 2018, doi: 10.1016/j.otorri.2017.05.006.
[51] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis—Jitter, shimmer and HNR parameters," *Proc. Technol.*, vol. 9, pp. 1112–1122, Jan. 2013. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2212017313002788
[52] S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox, "Formant centralization ratio: A proposal for a new acoustic measure of dysarthric speech," *J. Speech, Lang., Hearing Res.*, vol. 53, no. 1, pp. 114–125, Feb. 2010, doi: 10.1044/1092-4388(2009/08-0184).
[53] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[54] R. S. Olson, R. J. Urbanowicz, P. C. Andrews, N. A. Lavender, L. C. Kidd, and J. H. Moore, "Automating biomedical data science through tree-based pipeline optimization," in *Proc. Eur. Conf. Appl. Evol. Comput.* in Applications of Evolutionary Computation, Berlin, Germany: Springer, 2016, pp. 123–137, doi: 10.1007/978-3-319-31204-0_9.
[55] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and D. Uchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html
[56] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *Proc. 20th Int. Conf. Pattern Recognit.*, 2010, pp. 3121–3124. [Online]. Available: https://ieeexplore.ieee.org/document/5597285

PEDRO MATIAS received the M.Sc. degree in biomedical engineering from the NOVA University School of Science and Technology, in 2021. He is currently working as Scientist at the Intelligent Systems Group of Fraunhofer AICOS. He collaborates in projects, where the creation of intelligence from the available data can provide valuable solutions in response to their main challenges. His main research interests include exploring machine learning and deep learning techniques for automatic knowledge extraction from time series, trying to bridge research and development areas to tackle real–world problems.

JOÃO COSTA graduated the M.Sc. degree in bioengineering, focused on the field of biomedical engineering from the Faculty of Engineering, University of Porto, in 2019.

He was awarded a Summer Internship at Fraunhofer Portugal AICOS, after winning a machine learning challenge regarding the identification of subjects using inertial sensor data obtained with smartphones. During his internship, he was involved in the development of paddle detection algorithms for surfers. He also developed his master's dissertation in the same institute, focused on the automatic application and analysis of Verbal Fluency Tests for early detection of Neurocognitive Disorders on elderly people. He is currently a Scientist at Fraunhofer AICOS, Portugal, focused on the application of natural language processing tools for the automatic processing of clinical text, and on signal analysis of pathological speech and respiratory sounds.

ANDRÉ V. CARREIRO received the Ph.D. degree in biomedical engineering from Técnico Lisboa–University of Lisbon, He is a Senior Researcher at the Intelligent Systems Group, Fraunhofer AICOS, Portugal. He has applied machine learning techniques, since 2010, and has been working with deep learning methods, since 2016, both in academia and industry, resulting in a balance between innovation and making sure such techniques are applied efficiently to solve real–world problems, in areas from healthcare to security and industry.

HUGO GAMBOA (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the Instituto Superior Técnico of University of Lisbon (IST UL), in 2007. He is a Co-Founder and the President of PLUX, a company that develops bio-signals monitoring wearable technology. He is currently a Researcher at the Laboratory for Instrumentation, Biomedical Engineering and Radiation Physics (LIBPhys), Faculdade de Ciências e Tecnologia of NOVA University of Lisbon (FCT NOVA), where he is also an Associate Professor at the Physics Department. Since 2014, he has been a Senior Researcher at the Fraunhofer Center for Assistive Information and Communication Solutions (AICOS). His research interests include bio-signals processing and instrumentation.

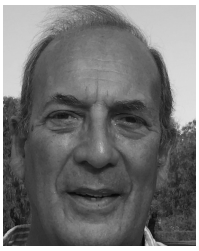
INÊS SOUSA received the Ph.D. degree in biomedical engineering from the University of Lisbon. Her thesis on Functional Magnetic Resonance Imaging of the Brain Using Quantitative Methods was developed at Siemens Healthcare, Portugal, and distinguished with the António Xavier Award.

She has been a Visiting Researcher at the Biomedical Imaging Center (CIBM), Ecole Polytechnique Fédérale de Lausanne, an Invited Lecturer at the Master's Program in radiations applied to health technologies, and a Research Intern at the Philips Research Europe. She is the Head of Intelligent Systems and a Senior Scientist at Fraunhofer, Portugal. She has been researching on machine learning, signal processing, and time series analysis, specifically focusing on human motion analysis based on inertial sensors. In the Intelligent Systems Department, she leads a team of around 25 researchers, who investigate ways in which artificial intelligence can augment human potential in interpreting vast amounts of data, identifying patterns, anticipating events, and supporting decisions.

Dr. Sousa has served as the principal investigator or a technical leader of several research projects in which she has collaborated with other researchers, healthcare providers, and companies to produce high quality functional prototypes and peer-reviewed publications. Her research interests include decision support systems, embedded and intelligent systems, and software engineering.

PEDRO GÓMEZ (Life Member, IEEE) received the M.Sc. degree in communications engineering and the Ph.D. degree in computer science from the Universidad Politécnica de Madrid, Madrid, Spain, in 1978 and 1983, respectively. He is a Professor at the Computer Science and Engineering Department, Universidad Politécnica de Madrid, since 1988. He is currently the Head of the Neuromorphic Speech Processing Laboratory, Center for Biomedical Technology. His current research interests include biomedical signal processing, neurological disease detection and monitoring from speech and voice, cognitive speech processing, and speaker biometry.

JOANA SOUSA received the M.Sc. degree in biomedical engineering with expertise in the areas of biosignal processing, innovation management, and marketing strategy. She is currently an Innovation Consultant at NOS Innovation, a Portuguese Telecommunications Company, and an independent expert in the European Community's H2020 framework, which provides funding for research and innovation. She was an Innovation Sub-Director at Edge Innovation, leading innovative projects and managing innovation in the Health, IT, communications areas, and a Research and Development Project Coordinator at PLUX–Wireless Biosignals, AS.



NUNO NEUPARTH graduated in medicine from the NOVA Medical School, in 1981. He received the master's degree in pedagogical skills and scientific capacity and the Ph.D. degree in medicine from the NOVA Medical School, in 1987 and 1995, respectively. He was aggregate in medicine, in 2017. He is a Full Professor of Pathophysiology at NOVA Medical School. He is a Vice-Dean at the NOVA Medical School and the President at the Pedagogical Council, since 2022. He was an Elected Member at the NOVA Medical School Council (2010–2021) and the NOVA Medical School Scientific Council, since 2010. He was also an Assistant (Specialist in Immunolaergology) at the Centro Hospitalar de Lisboa Central EPE, a Nominee Member at the Scientific Council of the National Program for Respiratory Diseases, the Director General of Health at the Ministry of Health, a Regent at the Curricular Unit of Pathophysiology and Therapeutic Targets I, NOVA Medical School. He was responsible for the Pathophysiology Course with the Integrated Master's degree in biomedical engineering at NOVA School of Sciences and Technology. He was also a Clinical Director at the Pathophysiology Laboratory, NMS/FCM. He was responsible for the Lung Function Laboratory at Dona Estefânia Hospital, Centro Hospitalar Universitário de Lisboa Central EPE (CHULC, EPE). He published 71 articles in specialized medical journals. He has 24 book chapters and six books. He organized 21 events. He supervised four doctoral theses and co-supervised four. He supervised seven master's dissertations and co-supervised three. He received 22 awards and/or honors. He participates and/or participated as an Investigator in five projects and a responsible investigator in 16 projects. His research interests include clinical medicine with an emphasis on allergology, medical and health sciences with an emphasis on health sciences, and public health and environmental health.

PEDRO CARREIRO-MARTINS received the Postdoctoral degree. He is an Associate Professor with Habilitation at the NMS−Faculdade de Ciências Médicas de Lisboa, where he coordinates the allergy course of the integrated master's programme and the pathophysiology course of the licentiate programme in nutritional sciences. He also works as an Allergy Consultant at the Centro Hospitalar Universitário Lisboa Central (Lisbon, Portugal), in one of the most important Portuguese Allergy Departments. He is a Researcher at the Comprehensive Health Research Center (CHRC), Nova Medical School. In recent years, he has been involved in the design and implementation of chronic respiratory disease projects and collaborated on m-health validation studies. He has coauthored over 70 journal articles. He has more than 1200 citations of articles with an H-index of 20 (Scopus).

He is the Vice-President of the Portuguese Society of Allergy and Clinical Immunology. He is a member of the Portuguese Allergy Committee (National Association of Medical Specialists). He is the Editor of the *European Annals of Allergy and Clinical Immunology*.

FILIPE SOARES received the Ph.D. degree (Hons.) in computer science and engineering from the University of Beira Interior, Portugal, in 2014. He is a Senior Scientist at the Fraunhofer Portugal Research Center for Assistive Information and Communication Solutions (AICOS), conducting research on computer vision and decision support systems in the field of retail, manufacturing, precision agriculture, and mHealth, in particular in the area of ophthalmology. He worked as a Researcher in medical imaging at Siemens Healthcare, in the field of computer-aided detection and diagnosis of breast cancer, in a project for which he received the Siemens Innovation Award, in 2012. He worked as a Software Engineer in telecommunications and embedded systems at Coriant (former Nokia Siemens Networks), in 2013. His main research interests include image processing, computer vision, machine learning, deep learning, decision support systems, embedded and intelligent systems, and software engineering.