
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Knowledge-Based Systems 253 (2022) 109539

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

# Deep learning and machine learning-based voice analysis for the detection of COVID-19: A proposal and comparison of architectures

Giovanni Costantini [a], Valerio Cesarini Dr. [a,*], Carlo Robotti [b,c], Marco Benazzo [b,c], Filomena Pietrantonio [d], Stefano Di Girolamo [e], Antonio Pisani [f,g], Pietro Canzi [b], Simone Mauramati [b], Giulia Bertino [b], Irene Cassaniti [h], Fausto Baldanti [c,h], Giovanni Saggio [a]

[a] Department of Electronic Engineering, University of Rome Tor Vergata, Rome, Italy
[b] Department of Otolaryngology — Head and Neck Surgery, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy
[c] Department of Clinical, Surgical, Diagnostic and Pediatric Sciences, University of Pavia, Pavia, Italy
[d] Internal Medicine Unit, Ospedale dei Castelli ASL Roma 6, Ariccia, Italy
[e] Department of Otorhinolaryngology, University of Rome Tor Vergata, Rome, Italy
[f] Department of Brain and Behavioral Sciences, University of Pavia, Pavia, Italy
[g] IRCCS Mondino Foundation, Pavia, Italy
[h] Molecular Virology Unit, Microbiology and Virology Department, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy

ARTICLE INFO

*Article history:*
Received 3 December 2021
Received in revised form 18 June 2022
Accepted 22 July 2022
Available online 28 July 2022

*Keywords:*
COVID-19
Speech processing
Classification
Deep learning
Adaboost

ABSTRACT

Alongside the currently used nasal swab testing, the COVID-19 pandemic situation would gain noticeable advantages from low-cost tests that are available at any-time, anywhere, at a large-scale, and with real time answers. A novel approach for COVID-19 assessment is adopted here, discriminating negative subjects versus positive or recovered subjects. The scope is to identify potential discriminating features, highlight mid and short-term effects of COVID on the voice and compare two custom algorithms. A pool of 310 subjects took part in the study; recordings were collected in a low-noise, controlled setting employing three different vocal tasks. Binary classifications followed, using two different custom algorithms. The first was based on the coupling of boosting and bagging, with an AdaBoost classifier using Random Forest learners. A feature selection process was employed for the training, identifying a subset of features acting as clinically relevant biomarkers. The other approach was centered on two custom CNN architectures applied to mel-Spectrograms, with a custom knowledge-based data augmentation. Performances, evaluated on an independent test set, were comparable: Adaboost and CNN differentiated COVID-19 positive from negative with accuracies of 100% and 95% respectively, and recovered from negative individuals with accuracies of 86.1% and 75% respectively. This study highlights the possibility to identify COVID-19 positive subjects, foreseeing a tool for on-site screening, while also considering recovered subjects and the effects of COVID-19 on the voice. The two proposed novel architectures allow for the identification of biomarkers and demonstrate the ongoing relevance of traditional ML versus deep learning in speech analysis.



## 1. Introduction

Voice production is a human skill relying on complex interactions between multiple systems, including air sources (lungs), a vibration mechanism (vocal folds), and resonant cavities (nasal, oral, pharyngeal and cranial). Such a skill is controlled by the brain and influenced by multiple surrounding factors, such as global health conditions, hydration and body temperature. Consequently, vocal samples may hold high informative content, since voice modifications may reflect the status of all the mentioned components. With these regard, distinctive vocal alterations have been identified and studied in various pathologies, mainly using

*Abbreviations:* ML, Machine Learning; CNN, Convolutional Neural Network; DL, Deep Learning; MFCC, Mel-frequency Cepstral Coefficients; P, Positive subjects; R, Recovered subjects; H, Healthy control subjects; NS, Nasal Swab; PCR, Polymerase Chain Reaction-based molecular swabs; 1E, Vowel /e/ vocal task; 2S, Sentence vocal task; 3C, Cough vocal task; PvsH, Positive versus Healthy subjects comparison; RvsH, Recovered versus Healthy subjects comparison; SVM, Support Vector Machine; CFS, Correlation-based Feature Selection; RF, Random Forest; ReLu, Rectified Linear Unit; ROC, Receiver-Operating Curve

* Correspondence to: Department of Electronic Engineering, Roma Tor Vergata University, Via del Politecnico 1, 00133 Rome, Italy.
*E-mail address:* valerio.cesarini@uniroma2.it (V. Cesarini Dr.).

https://doi.org/10.1016/j.knosys.2022.109539
0950-7051/

machine learning (ML)-based methods, yielding encouraging outcomes. Understandably, research focused extensively on primary affections of the phonatory apparatus: specifically, Suppa et al. [1] investigated essential tremor, Teixeira et al. [2] assessed chronic laryngitis, Costa et al. [3] investigated vocal fold edema, Petrovic-Lazic et al. [4] revealed vocal polyps, and Alves et al. [5] reviewed the changes in voice quality related to hydration conditions and Zacharia et al. [6] explored head and neck cancer. Interestingly, promising results were also highlighted for pathologies leading to vocal alterations only secondarily, such as neurodegenerative pathologies (Parkinson's, SLA) [7,8], Down syndrome [9], and even cardiovascular disorders [10–12] as well as movement disorders [13].

That being said, the COVID-19 pandemic represents an ideal and urgent field of application for this line of research. As a matter of fact, since SARS-CoV-2 can affect both the respiratory apparatus and the nervous system [14], it is possible to surmise that the disease may likely alter the sound of both voice and cough. Therefore, technologies able to detect vocal biomarkers of this infection could provide national health authorities with additional and non-invasive surveillance strategies.

Deep learning (DL), especially based on Convolutional Neural Networks (CNN) applied to spectral images, is commonly considered as the major alternative to traditional ML pipeline methods for speech assessment. Indeed, it does offer advantages which include the extraction of very complex features – through repeated non-linear data transformation – and its completely data-driven nature, which allows to forgo data pre-processing. However, DL is also considered to perform unsatisfactorily on small datasets, requiring larger ones by its very nature [15], preferably in association with data augmentation procedures that inherently bring a degree of knowledge-based processing. Furthermore, DL are computational heavy and based on a large number of high-level parameters. Examples are studies by Sztahó et al. [16] and Nissar et al. [17], which highlighted the possibilities of DL for pathological speech assessment, although datasets and accuracies are comparable to those obtained using ML methods for the same pathologies, namely Parkinson's disease [18] and dysphonia [19].

As for COVID-19, Laguarta [20] applied a CNN model on crowdsourced cough sounds, recognizing patients with a 97% accuracy. Similarly, Imran et al. [21] obtained an accuracy higher than 90% through principal component analysis (PCA) and data processing using mel-frequency cepstral coefficients (MFCC). The algorithms presented by Pinkas [22] and Shimon [23] yielded average accuracies around 83%, while Despotovic [24] achieved 88% in accuracy considering vocal, speech, cough and breathing crowdsourced sounds. To the best of our knowledge, the only study involving recovered subjects is the one by Suppakitjanusant et al. [25], using CNN with a mean accuracy of 74%. More recently, the Interspeech 2021 conference proposed the DiCOVA challenge, with teams testing algorithms for COVID-19 detection on crowdsourced voice samples [26], with a baseline mean accuracy of 73% and the highest one being 87%.

All in all, we consider DL and traditional ML to be equally powerful and their usage to be significantly problem-dependent, so that here we employ fine-tuned versions of both to compare the results they can furnish. The chosen vocal tasks consisted in a sentence, a sustained vowel, and solicited coughing, so to gather somehow different informative content. The study population was comprised of three matched groups: COVID-19 positive patients, COVID-19 recovered individuals, and healthy subjects as controls.

We adopted a multifaceted approach based on state-of-the-art ML algorithms, training an AdaBoost-based algorithm and a CNN-based one independently, albeit SVM classifiers [27] were considered as well. With these regards, we consider our innovation to be in the construction of a relatively homogeneous, polished dataset, with suitable domain-specific pre-processing and the usage of custom ML algorithm, which in turn detail and expand the existing state-of-the-art of voice analysis. In addition to the high accuracy and sensitivity results, we found acoustic vocal biomarkers for the identification and study of COVID-19, also taking into account the staging of the disease and its recovery, and began foreseeing a potential automatic tool for the real-time remote pre-screening. More on this will be detailed in the Discussion section.

## 2. Materials

### 2.1. Study population

Three groups of subjects were enrolled in the present study, namely #70 COVID-19 positive patients (group P, or P for short), #120 recovered COVID-19 individuals (R) who were initially proven positive and then negative, and #120 healthy control subjects (H) who never got infected. All of them were of Caucasian ethnicity. Average age and gender were: 57 yo (range 39–67), 57% male, for P subjects; 53 yo (range 39–69), 52% male, for R subjects; 50 yo (range 29–57), 54% male, for H subjects. Informed consent was obtained from all participants and all data was pseudonymized. Patient enrollment was carried out at three different Italian institutions: "San Matteo" University Hospital in Pavia (Otolaryngology Unit, ethical approval number 20200053388), "Tor Vergata" University Hospital in Rome (Otolaryngology Unit, ethical approval number 0012909/2020), and the "Dei Castelli" Hospital in Rome (General Medicine Unit, ethical approval number 0064181/2020).

Patients of group P were recruited within ten days from nasal swab (NS) positivity (RT-PCR), and COVID-19 pneumonia was diagnosed clinically and radiologically through a chest computed tomography (CT) scan. Subjects of group R were initially tested positive via RT-PCR NS and subsequently proved negative with two consecutive swabs. Finally, subjects of group H, recruited among hospital staff members and their acquaintances, had no COVID-19 symptoms, nor reported unprotected exposure to confirmed or suspected COVID-19 cases, with their serum samples, collected at least 20 days after the vocal tests, yielding negative results for both IgG and IgM antibodies.

To guarantee as homogeneous as possible a dataset for voice analysis, in addition to the average age and gender distribution being approximately matching among the three groups, data regarding clinical and demographic characteristics of the study population was also collected. Non-smokers represented almost half of participants in each study group (51% for group P, 54% for group R, 52% for group H). As far as clinical features are concerned, one or more COVID-19 symptoms (muscle pain, dyspnea, asthenia) were present in 78% and 75% of P and R subjects, respectively. Conversely, at the time of recording, cough was reported by 49% and 8% of P and R subjects, respectively. Finally, in order to minimize the heterogeneity of the dataset, more polarizing characteristics which could greatly affect breathing, articulation or voice emissions (like C-PAP therapy) were deemed as exclusion criteria. Table 1 depicts the main inclusion and exclusion criteria for all groups.

### 2.2. Vocal tasks and recordings

We considered different vocal tasks to gather heterogeneous informative content, namely: the sustained vowel/e/ (1E); the popular Italian proverb *"A caval donato non si guarda in bocca"* (2S) and solicited cough (3C).

The vowel sound involves a quasi-periodic vibration of the vocal folds. Furthermore,/e/ is produced keeping the larynx in an

**Table 1**
Inclusion and exclusion criteria.

| Inclusion criteria | P | R | H | Exclusion criteria | P | R | H |
|---|---|---|---|---|---|---|---|
| 18–80 yo age range | ✔ | ✔ | ✔ | Drugs acting on CNS | ✔ | ✔ | ✔ |
| European ethnicity | ✔ | ✔ | ✔ | Head/neck cancer | ✔ | ✔ | ✔ |
| Italian native speaker | ✔ | ✔ | ✔ | Lung cancer | ✔ | ✔ | ✔ |
| Positive NS (< 10 days) | NA | ✔ | NA | Chemoradiation therapy | ✔ | ✔ | ✔ |
| Two consecutive negative NS | NA | ✔ | NA | C-PAP Therapy | ✔ | ✔ | ✔ |
| LUS $\leq$ 3 | NA | ✔ | NA | Tracheal intubation | ✔ | ✔ | ✔ |
| Negative SS test (< 20 days) | NA | NA | ✔ | Tracheostomy | ✔ | ✔ | ✔ |

Abbreviations: NS: SARS-CoV-2 nasal swab for RNA detection; LUS: lung ultrasound score; SS: SARS-CoV-2 serum sample for IgM and IgG quantification; CNS: Central Nervous System; C-PAP: Continuous Positive Airway Pressure; LUS: Lung Ultrasound score; NA: not applicable.

almost neutral position, therefore avoiding artifacts due to excessive muscular contraction [28], while still being able to reflect possible pathological alterations of the lower respiratory tract. The sentence was necessary to study the vocal characteristics of speech (including prosody) and their deviations. Furthermore, the selected saying holds a prevalence of plosive consonants, the phonation of which is associated to the production and explosive emission of a relevant amount of air. Finally, cough sounds were selected since they can be reflective of possible alterations of the lungs and of the lower respiratory tract as a whole [20].

One participant, at turn, was comfortably seated, with arms resting on the armrests at the center of the room, and was asked to perform each vocal task twice, to select the best ones (one for each task) in terms of noise and intelligibility. Each participant was asked to sustain the vowel steadily for 2 to 5 s without straining, then to pronounce the sentence without pausing between words and at a natural speaking tone, and finally to cough for three consecutive times.

The recordings were captured through a smartphone (Y6s, by Huawei Technologies Co., Ltd., Shenzhen, China), kept at about 20 cm from the mouth, with the aid of a web-app (https://covid19.voicewise.it). Audio was recorded in .wav format, sampled at 44.1 kHz, with a resolution of 16-bit. We opted for a smartphone so that it could become an easily adoptable solution for a worldwide and low-cost adoption, as detailed in the Discussion. Recording sessions were held in rooms which were comparable in terms of acoustics and dimensions and had an appropriately quite environment (low-reverberation and low-noise levels). Recordings were only accepted when no hiss nor hum noises were detectable; additionally, no machines nor background voices were captured. The overall audio quality (background noise, reverberation, intelligibility) was subsequently assessed by ear by independent audio engineers.

The vowel and sentence audio files were trimmed so to remove noises and silence at the beginning or the end. The cough files, originally comprehending three coughs in one recording, were split into three different files to isolate each single cough sound.

Trimming was performed automatically with custom-made routines in MATLAB (by Mathworks Inc., Natick, Massachusetts, USA [29]) based on compression and expansion followed by a "lowess" (locally weighted scatterplot) smoothing [30], band-pass filtering, and thresholds based on RMS Energy and MFCC. All recordings were also examined by sound experts to check the correct trimming; manual corrections were eventually applied when necessary using audio editing solutions available within the digital audio workstation REAPER (Cockos Inc., San Francisco, California, USA).

### 2.3. Datasets for classification

Since three different vocal tasks were considered, three binary classifications were necessary for each comparison. For the comparison between groups P and H (PvsH), in order to have the same number of instances in each class, we down-sampled 70 out of the 120 subjects of group H with age-range and gender distributions similar to those of group P. Consequently, the PvsH comparison was based on datasets of 70 subjects per task. Therefore, 70 instances for both the tasks 1E and 2S, as well as 210 instances for the task 3C, were analyzed for each group.

For the comparison between groups R and H (RvsH), the final datasets consisted of 120 subjects in total. That translated to 120 instances for tasks 1E and 2S, and 360 instances for the task 3C.

Finally, the total dataset was split randomly as follows: 85% as the training set, 15% as the validation set, which was never fed to any algorithm. Since the task 3C originally involved three instances per subject, only one of those was retained in the validation set. The other two were left out of the validation set as they would have been redundant, being representative of the same subject. For each binary classification (PvsH and RvsH) two different approaches were explored (Adaboost and CNN), each one encompassing three sub-classifiers, one for each speech task (1E, 2S, 3C).

## 3. Methods

The following paragraphs describe the two different approaches used for both classifications: a Random Forest-based Adaboost approach applied on selected acoustic features, and two CNN architectures applied to augmented mel-spectrograms.

### 3.1. Adaboost approach

The Adaboost classifier-based approach is divided into four steps for each sub-classifier, as follows:

(1) Audio feature extraction, with features reported into a single data matrix with two classes;
(2) Feature selection, using a Correlation-based Feature Selection algorithm (CFS) with a Forward Greedy Stepwise search method;
(3) Additional feature selection, with a wrapper-based ranker embedding a linear SVM classifier;
(4) Training of an AdaBoost classifier with Random Forest weak learners.

All the Machine-Learning processes (steps 2 to 4) were implemented within the Weka platform (University of Waikato, New Zealand, GNU General Public License [31]), while the feature extraction was performed using OpenSMILE (Audeering GmbH, Gilching, Germany [32]).

#### 3.1.1. Feature extraction

The feature set we chose had a total of 6373 acoustic features defined within the INTERSPEECH2016 Computational Paralinguistics Challenge (ComParE) [33], carrying several computational functionals (e.g. mean, position of peak, quartiles, delta coefficients) in the time, spectrum and cepstrum domains [34]. Some

of the relevant features include the relative spectral (RASTA) PLP coefficients [35], the voicing probability [36], and the spectral loudness summation [37].

### 3.1.2. Feature selection

To address the discrepancy between the number of features and the number of training instances, we performed a feature reduction, according to [38,39], using a correlation-based feature selection (CFS) algorithm. Specifically, this algorithm takes into account the redundancy of the features in a subset, and their eventual correlation with the class itself [40]. The basic principle is the computation of a merit factor for subsets of features, according to the equation:

$$M_S = \frac{k * r_{fc}}{\sqrt{k + k\,(k-1) * r_{ff}}}$$

where $k$ is the number of features in the subset $S$, $r_{fc}$ is the average correlation between each feature in the subset and the class, and $r_{ff}$ is the average cross-correlation between all the features one with each other.

A forward greedy stepwise search method, chosen as a good trade-off between computational time and exhaustiveness [41], allowed for the selection of the subset, resulting in a number of features which spans from 1% to 3% of the full 6373-features set. A further reduction was applied to the feature subsets, which was also useful to work with a homogeneous number of features throughout all sub-classifications. A wrapper-based feature selector employing a soft-margins linear SVM [42], trained with Platt's SMO Optimizer [43] on a single feature at a time, was used to perform ranking. We empirically considered the first 50 ranked features, and retained them as the final set to train the Adaboost classifier.

### 3.1.3. Classification

According to literature, SVM and tree-based classifiers – Random Forests (RF) in particular – are the most common solutions for audio-related classification, especially when it comes to speech analysis [44–47]. For the present study, we adopted RF as a boosted learner in reason of its strength when it comes to non-linear classification problems [48]. The AdaBoost M1 method was applied employing RF internal classifiers ("weak learners"). This approach was chosen for its proven effectiveness in voice classification [49]. In addition, basing on literature [50,51] and on our experience, boosting proved beneficial to voice analysis when using more complex weak learners.

Adaboost M1 is a "boosting" technique aimed at generalizing ensembles of weak learners by running on various weight distributions over the training data, finally combining the obtained classifiers into one [52]. Weights are updated so that training examples which are difficult to classify get assigned a higher weight. Ultimately, among all the possible alternatives, the final predicted label is the one which maximizes the sum of the (logarithmic) reciprocals of the prediction error. In particular, the predicted label of the RF classifier is decided by majority voting over simple decision trees trained on different subsets of the training set ("bags" sampled with repetition) and features [53,54].

AdaBoost with RF has proven effective especially for problems where RF itself represents a suitable solution, providing further improvement in error rates, mostly in case of non-linear dependencies and dataset-related complexities [51,55]. Although both bagging and boosting aim at producing a low variance hypothesis combining higher variance ones, they actually succeed in the task in different ways: the former creates subsets of the training data and works in parallel on them, while the latter manages the training space as a whole, repeatedly weighing it with different distributions.

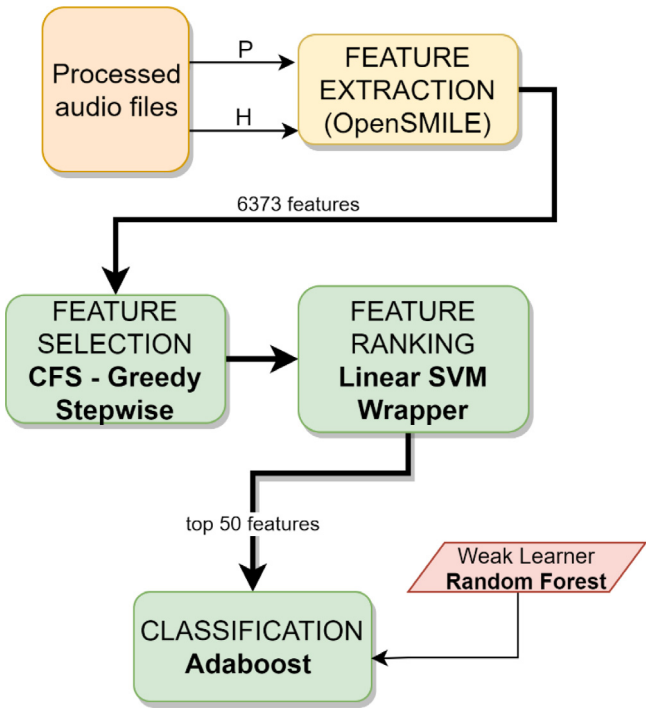


**Fig. 1.** Flowchart describing the complete pipeline of the Machine Learning approach based on the Adaboost classifier (exemplified for the PvsH comparison).

For the present study, in order to emphasize the different dynamics of boosting and bagging, each bag for the RF bootstrapping was built using 80% of the whole training set on each bag. A number of 1000 iterations was selected for both AdaBoost and the internal RF.

The main steps for the whole process, exemplified for the PvsH comparison, are displayed in Fig. 1. The same process was also carried out for the RvsH comparison.

## 3.2. Convolutional Neural Networks (CNN) approach

Convolutional Neural Networks (CNN) and DL have gradually become gold standards for voice analysis [56]. Most problems regarding voice analysis for health-related issues involve CNN at some point, and COVID-19 studies are no exception [20,57]. As a matter of fact, most of the studies described in the Introduction achieved the best results either with CNN or RF classifiers. However, as suggested by Cummins et al. [58], DL is still unable to outclass the efficacy of traditional ML in voice analysis for multiple limitations, such as the complexity of acoustic features and the scarcity of datasets.

### 3.2.1. Mel-spectrograms as image inputs

CNN are mostly employed on images in reason of their filtering nature. Indeed, they effectively identify local graphical features, to the point that the logical process behind CNN may somehow recall human sight [59]. Therefore, even for audio applications, graphic plots are preferred as inputs. Mel-spectrograms of all audio recording were therefore produced and exported as grayscale .png images. Subsequently, 4096 points FFT mel-spectrograms were generated with a 50% overlapped Hamming window. Since invaluable information for vocal tasks is found in frequency bands extending up to a few kHz [60], the frequency range for the spectrograms was limited to 20–12000 Hz, also limiting the complexity of the problem. These steps, as well as the whole CNN procedure, was implemented in MATLAB.

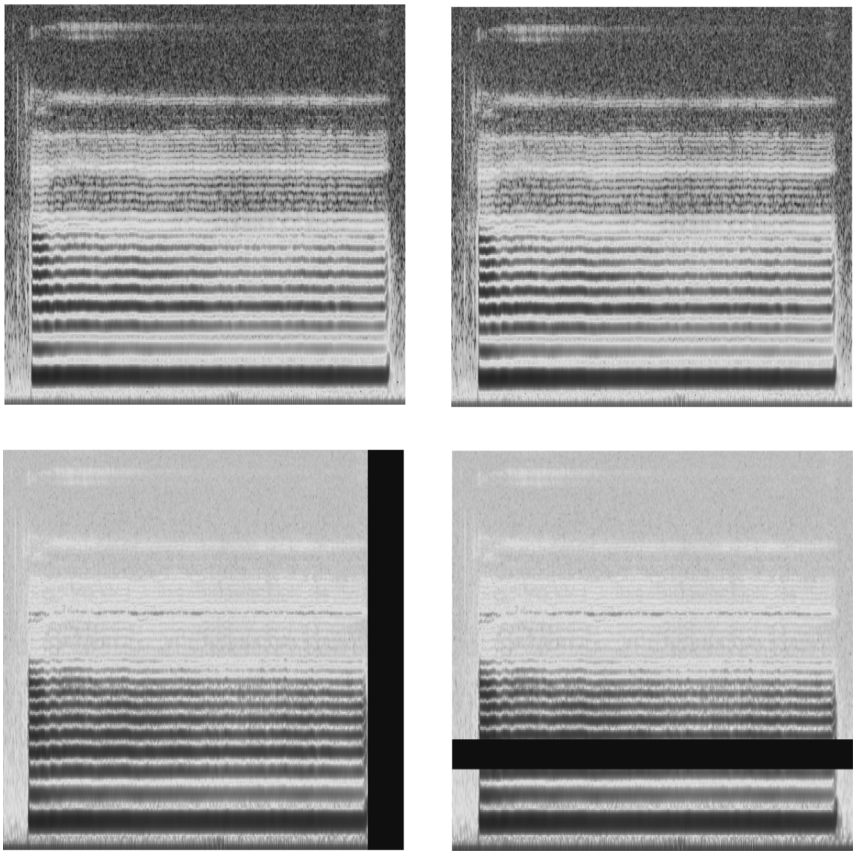

**Fig. 2.** Visualization of the data augmentation techniques applied to mel-frequency spectrograms.
Top left: original sample spectrogram, top right: pink noise addition, bottom left: time masking, bottom right: frequency masking. .

*3.2.2. Data augmentation*

Since one of the main drawbacks of DL is the requirement for a large pool of data [61], augmentation procedures have become common practice to improve the generalization and the accuracy of models [62]. Successful results in augmenting biometric data for the detection of COVID-19 were obtained by Barshooi and Amirkani [63] with a novel approach based on synthetic GAN-generated data, pre-processed by a Gabor filter. However, the images fed to our CNN are in fact time plots, and most graphical artifacts and/or synthesis methods would result in unrealistic augmented data which would bring in the risk of biasing the net.

All the employed augmentations were either a modeling of a real-world audio effect, or appropriate mathematical/graphical artifacts on the spectrograms. Specifically, the selected data augmentation methods were:

(1) Pink noise addition;
(2) Frequency masking on the spectrogram;
(3) Time masking on the spectrogram;
(4) Frequency and time masking used together.

Pink noise was preferred over white noise for multiple reasons, such as the invariance of its energy content with respect to pitch and its resemblance to real-world noise able to affect the data [64]. Frequency and time masking are artifacts proven effective by Park in the SpecAugment study [65]: they involve the "masking" of a random range of frequency or time on the spectrogram, which are represented by a dark horizontal and vertical band, respectively. All data augmentation methods, whose results on the spectrogram images can be seen in Fig. 2, were implemented using random values for both the starting point and the width of the masked bands, for which the values of the spectrogram were brought very close to zero ($10^{-15}$). A SNR of 35 dB was chosen for pink noise addition, whose signal was randomly generated and then added on MATLAB. Audio files were normalized again after the addition of noise to avoid clipping.

Data augmentation was carried out only on the training set, and it resulted in new training sets five times larger than the original ones. Specifically, for the PvsH comparison, data augmentation produced 300 instances for 1E and 2S, and 900 instances for 3C. For the RvsH comparison, the process led to 510 instances for 1E and 2S, and 1530 instances for 3C.

*3.2.3. Proposed architectures*

Although transfer learning proved effective in voice analysis [66], custom-built and simpler architectures were preferred for the present research project. This choice was mainly dictated by our knowledge of the complexity of audio classification tasks, as well as by the chance of a faster and more controllable framework for future implementations of the tool. However, we did also experiment on transfer-training several popular CNNs (namely AlexNet and ResNet50), finding no improvements over our custom architectures.

Two different CNN architectures were built: one (CNN1) was employed for the two "vocalized" tasks (1E and 2S), the other (CNN2) for the cough task (3C). These architectures were used in both comparisons (PvsH and RvsH). Square, grayscale images ($257 \times 257$ pixels) were used as inputs. As depicted in Fig. 3, CNN1 encompasses a total of 6 convolutional layers with a growing number of $3 \times 3$ filters, from 16 to 128. Each convolutional layer is followed by a batch normalization layer and a ReLu activation function. This ensemble is represented as "Conv Block" in the figure. Max Pooling layers are $2 \times 2$ in size with a Stride of 2. An additional fully connected (FC) layer containing 128 neurons,

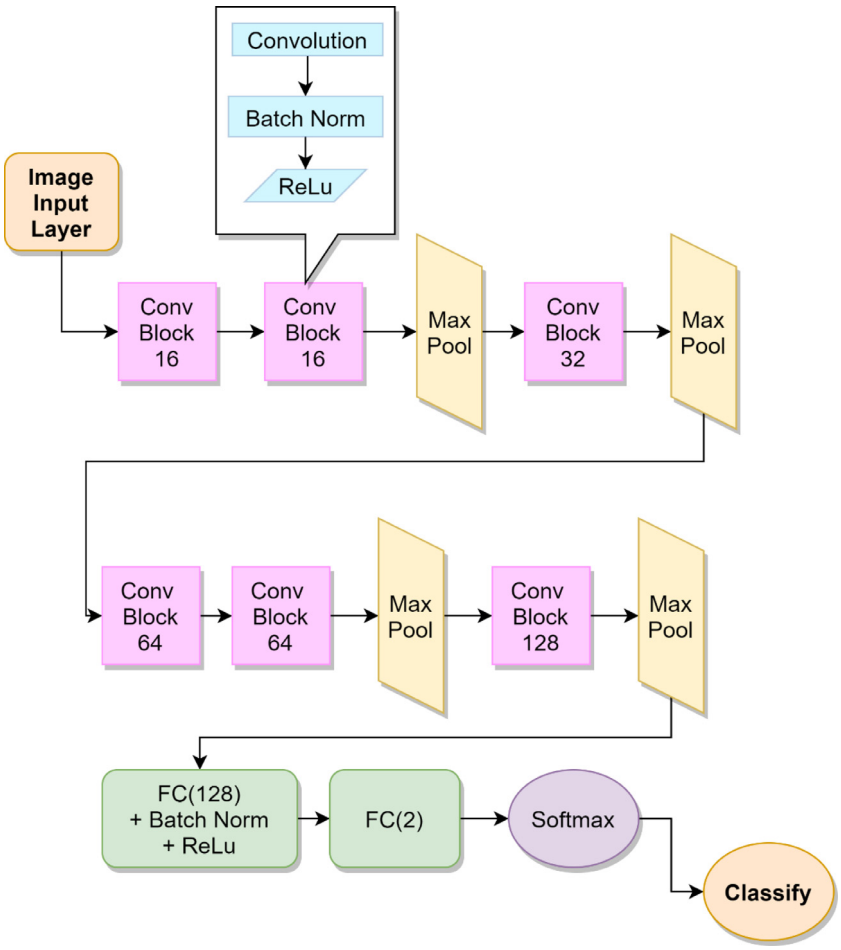


**Fig. 3.** CNN1 architecture.
A "Conv Block" is described in the higher box and is comprised of a convolutional layer followed by a batch normalization layer and a ReLu (Rectified Linear Unit) activation function. The number after "Conv Block" indicates the number of parallel convolutional filters/neurons in the layer. Max Pool: max pooling layer; FC: Fully connected layer: the number in the round brackets indicates the number of neurons.

followed by batch Normalization and ReLu, precedes the last FC layer before the classification output.

CNN2, displayed in Fig. 4, is a simplification of CNN1 used for the task 3C, containing three convolutional layers with 16, 64 and 128 filters respectively, before the two fully connected layers.

For the training of both nets the ADAM Optimizer was used [67], with a gradient decay factor of 0.8, employing L2-Regularization [68] and a piecewise learning rate decaying with a factor of 0.8 every 10 epochs.

## 4. Results

### 4.1. Accuracies

For both classification approaches, a total of three sub-classifiers per comparison (PvsH and RvsH) was built, that is, one for each different vocal task. The final predictions on the validation set were unified by means of a majority voting method, considering each of the three sub-classifiers having the same weight. Thus, two or three errors on an instance lead to misclassification in the final result (each instance in the validation set corresponds to a different person).

For the PvsH comparison, instances 1 to 10 are of healthy subjects, while instances 11 to 20 correspond to COVID-19 positive ones. For the RvsH comparison, healthy subjects go from 1 to 18, while the remaining two are recovered ones.

Table 2 and Table 3 show the results of Adaboost and CNN respectively, with confusion matrices highlighting the errors in the single sub-classifiers and presenting the final output obtained by majority voting.

In order to better interpret the results, the concepts of sensitivity and specificity must be introduced, as they represent useful measures for binary classifications, especially in the biomedical field. Sensitivity – or true positive rate – is the ratio of positive subjects correctly identified as such ($TP$) versus all the positives ($P_{tot}$), following the formula $Sensitivity = TP/P_{tot}$. The specificity – or true negative rate – refers to the correctly identified negative subjects ($TN$) versus all the negatives ($N_{tot}$), according to $Specificity = TN/N_{tot}$. In the PvsH comparison, positive subjects are indeed the individuals with an ongoing COVID-19 infection; whereas in the RvsH comparison, the R class is considered as "positive" for the scope of calculating sensitivity and specificity. For the purposes of the present study, a high sensitivity was considered to be a priority objective, especially for the PvsH comparison.

As far as the PvsH comparison is concerned, an accuracy of 100% was obtained with the Adaboost approach; the CNN approach reached a 95% accuracy, which translates to one misclassification. It is also worth noting that the only misclassified subject is a healthy one, which means that the CNN-based approach also has a sensitivity of 100%. The RvsH comparison yielded less accurate results, with the AdaBoost-based approach reaching 86.1%

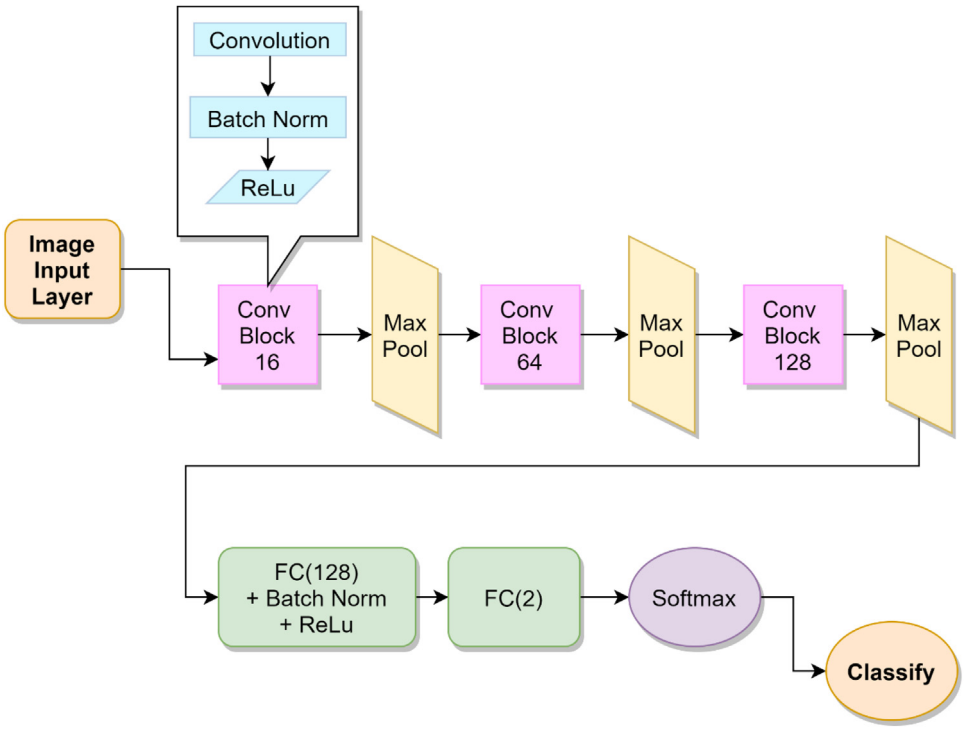


**Fig. 4.** CNN2 architecture (for the sole 3C — Cough vocal task).
A "Conv Block" is described in the higher box and is comprised of a convolutional layer followed by a batch normalization layer and a ReLu (Rectified Linear Unit) activation function. The number after "Conv Block" indicates the number of parallel convolutional filters/neurons in the layer. Max Pool: max pooling layer; FC: Fully connected layer: the number in the round brackets indicates the number of neurons. .

**Table 2**
Confusion matrices for the PvsH comparison over the two classification approaches (Adaboost and CNN).

| #Inst | Real class | Adaboost | | | | CNN | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1E | 2S | 3C | Final | 1E | 2S | 3C | Final |
| 1 | H | – | – | X | ok | – | – | – | ok |
| 2 | H | – | – | X | ok | – | – | – | ok |
| 3 | H | – | – | – | ok | – | – | – | ok |
| 4 | H | – | – | – | ok | – | – | – | ok |
| 5 | H | X | – | – | ok | X | X | – | X |
| 6 | H | – | – | – | ok | – | – | – | ok |
| 7 | H | – | – | – | ok | – | – | X | ok |
| 8 | H | – | – | – | ok | – | X | – | ok |
| 9 | H | – | – | – | ok | – | – | – | ok |
| 10 | H | – | X | – | ok | – | – | – | ok |
| 11 | P | – | – | – | ok | – | – | X | ok |
| 12 | P | – | X | – | ok | – | X | – | ok |
| 13 | P | – | – | X | ok | – | – | – | ok |
| 14 | P | – | – | – | ok | – | – | – | ok |
| 15 | P | – | – | X | ok | X | – | – | ok |
| 16 | P | – | X | – | ok | – | – | – | ok |
| 17 | P | – | – | X | ok | – | – | – | ok |
| 18 | P | – | X | – | ok | – | – | – | ok |
| 19 | P | – | – | – | ok | – | – | – | ok |
| 20 | P | – | – | – | ok | – | – | X | ok |
| **Accuracy (%)** | | 95 | 80 | 75 | **100** | 90 | 85 | 85 | **95** |

Abbreviations: #Inst: Number of test instance; H: Healthy group; P: Positive group; 1E: Sustained vowel /e/ vocal task sub-classifier; 2S: Sentence vocal task sub-classifier; 3C: Cough vocal task sub-classifier; CNN: Convolutional Neural Network approach; -: No error in sub-classifier; X: Classification error; Final: Final classification output obtained by means of majority voting of the three (1E, 2S, 3C) sub-classifiers; ok: No final classification error.

and the CNN reaching 75%. However, both classifiers interestingly yielded 100% sensitivity in identifying COVID-19 recovered patients.

### 4.2. Acoustic features for the adaboost approach

The top ranked features employed for the AdaBoost-based approach can be considered as clinical "biomarkers" for COVID-19 identification through voice analysis.

Most works based on similar vocal tasks are based on different acoustic features, including MFCC, HNR, jitter, shimmer and fundamental frequency (F0) [69,70]. The most frequent feature domains as assessed by the feature selection are reported in Table 4. According to our results, relevant features for determining the positivity to COVID-19 include the RASTA-PLP Coefficients, which ranked higher than MFCC. The RASTA-PLP processing can be considered somehow similar to MFCC [71], and it is especially aimed for speech signals due to its insensitiveness to slowly varying background noises [72]. In fact, this processing is especially sensitive to background voices, which we carefully and intentionally avoided. RASTA is widely used in speech recognition and, to the best of our knowledge, it has yet to be solidly introduced in studies regarding voice analysis for healthcare. This is also in line with other studies carried out by our study group on different diseases, especially regarding the implications of higher-numbered RASTA windows and "rough" voices [73].

Voicing Probability-related features appear to be the most relevant for the vowel task, while different frequency domain features are present in all the sets.

Receiver-operating curves (ROC) related to PvsH and RvsH for the AdaBoost classifiers are presented in Fig. 5. The area under the curve (AUC) values for the PvsH are 0.94 for task 1E (red), 0.90 for task 2S (blue) and 0.88 for task 3C (green), respectively. For the RvsH comparison they are 0.74 for 1E, 0.98 for 2S and 0.85 for 3C, respectively.

Fig. 6 shows a radar plot for an interesting, exemplified view of the differences among the acoustic features. The plot shows the average, over all the subjects, of the top 20 features. The features are normalized by the average of the H-group class, which is consequently represented by a unit circle.

**Table 3**
Confusion matrices for the RvsH comparison over the two classification approaches (Adaboost and CNN).

| #Inst | Real class | Adaboost | | | | CNN | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1E | 2S | 3C | Final | 1E | 2S | 3C | Final |
| 1 | H | X | – | X | X | X | – | X | X |
| 2 | H | X | – | – | ok | X | – | – | ok |
| 3 | H | X | – | – | ok | – | – | X | ok |
| 4 | H | X | – | X | X | X | – | X | X |
| 5 | H | X | – | – | ok | X | – | X | X |
| 6 | H | X | – | – | ok | X | – | – | ok |
| 7 | H | – | – | – | ok | – | X | X | X |
| 8 | H | – | – | X | ok | X | – | X | X |
| 9 | H | – | – | – | ok | X | – | X | X |
| 10 | H | – | – | – | ok | – | – | – | ok |
| 11 | H | – | – | – | ok | X | – | – | ok |
| 12 | H | – | – | – | ok | – | – | – | ok |
| 13 | H | – | X | X | X | X | – | X | X |
| 14 | H | – | – | X | ok | – | – | X | ok |
| 15 | H | X | X | – | X | X | X | X | X |
| 16 | H | X | X | – | X | X | – | – | ok |
| 17 | H | X | – | – | ok | – | – | – | ok |
| 18 | H | X | – | – | ok | X | X | – | X |
| 19 | R | – | – | X | ok | – | – | – | ok |
| 20 | R | – | – | – | ok | – | – | – | ok |
| 21 | R | – | – | – | ok | – | – | X | ok |
| 22 | R | X | – | – | ok | – | – | – | ok |
| 23 | R | – | – | – | ok | – | – | – | ok |
| 24 | R | X | – | – | ok | – | – | – | ok |
| 25 | R | – | – | – | ok | – | – | – | ok |
| 26 | R | – | X | – | ok | – | – | – | ok |
| 27 | R | – | – | – | ok | – | – | – | ok |
| 28 | R | – | – | X | ok | – | – | – | ok |
| 29 | R | – | – | – | ok | – | – | – | ok |
| 30 | R | – | – | X | ok | – | – | – | ok |
| 31 | R | – | – | – | ok | – | – | – | ok |
| 32 | R | – | – | – | ok | – | – | – | ok |
| 33 | R | – | – | – | ok | X | – | – | ok |
| 34 | R | – | – | X | ok | – | – | – | ok |
| 35 | R | – | – | X | ok | – | – | – | ok |
| 36 | R | – | – | – | ok | – | – | – | ok |
| **Accuracy (%)** | | 66.7 | 88.9 | 72.2 | **86.1** | 63.9 | 91.7 | 69.4 | **75.0** |

Abbreviations: #Inst: Number of test instance; H: Healthy group; R: Recovered group; 1E: Sustained vowel /e/ vocal task sub-classifier; 2S: Sentence vocal task sub-classifier; 3C: Cough vocal task sub-classifier; CNN: Convolutional Neural Network approach; -: No error in sub-classifier; X: Classification error; Final: Final classification output obtained by means of majority voting of the three (1E, 2S, 3C) sub-classifiers; ok: No final classification error.

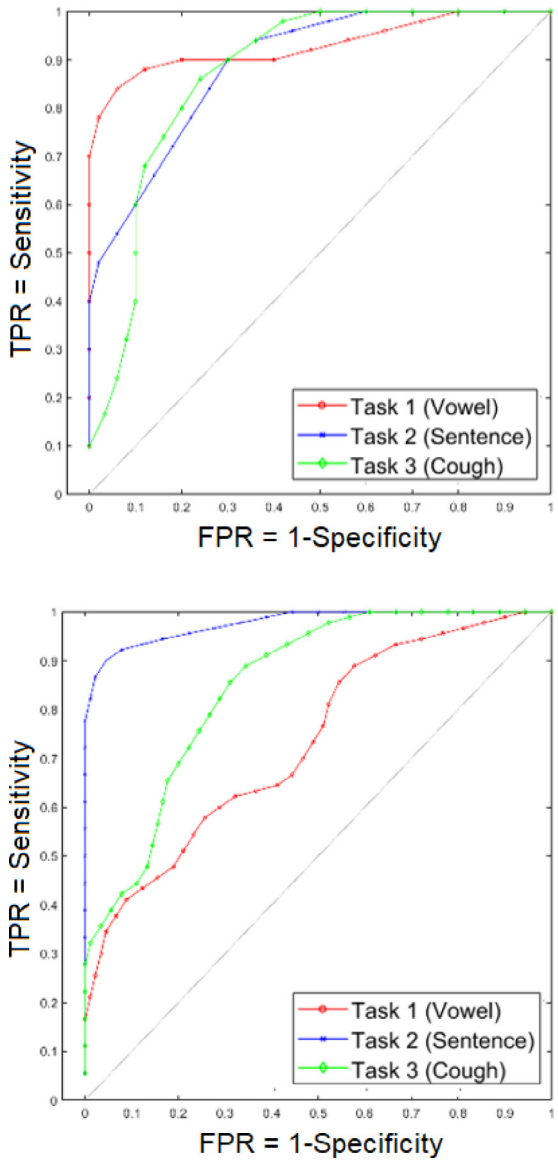


**Fig. 5.** ROC curves.
Above: ROC curve for the PvsH (Positive versus Healthy) comparison. Below: ROC curve for the RvsH (Recovered versus Healthy) comparison. Red line refers to the 1E — vowel/e/ vocal task sub-classifier; blue line refers to the 2S — sentence vocal task sub-classifier; green line refers to the 3C — cough vocal task sub-classifier. Axes span from 0 to 1. AUC (Area Under the Curve) values are reported in the manuscript.

## 5. Discussion

The human ear does not possess sufficient sensitivity to distinguish different pathological conditions just by listening to patients' voices, even though well-trained and experienced clinicians may sometimes obtain precious hints for diagnosis from perceptual (by-ear) voice analysis, in particular for pathologies with very peculiar features of voice and speech [74,75]. On the other hand, several studies demonstrate the possibility to identify vocal, respiratory and even neurological diseases from the automatic analysis of the speech signal.

At the present time, a fair amount of studies have tried to identify COVID-19 from the human voice. However, common issues in voice analysis are exacerbated by the difficult pandemic situation, which makes it very hard to gather a reasonable amount of high-quality data, as well as the short timespans, and the limited knowledge on the disease.

Our aim was to build a reliable framework based on a wide amount of information contained in high-quality data, with the best and most reproducible recording conditions that, working with subjects with proven clinical status. Since state-of-the-art methodologies contemplate both ML and CNN-based solutions, we chose to employ and to compare both with custom fine-tuned architectures and knowledge-based pre-processing on the input data. As of today, no reliable datasets exist for COVID-19 speech, which made the construction of an independent validation set one of the only possible choices.

Table 5 reports the main characteristics of our study compared to most of the other published works. Although we assert the experimental and preliminary nature of these studies, we believe that ours is the first one to achieve such promising results with the use of non-crowdsourced, verified audio data, while also considering recovered subjects, bringing novel architectures and comparing the most technologically advanced methods for speech analysis.

The chosen approaches exemplify the division between traditional ML pipelines allowing for a better control of every step of the inference, and CNNs, which act almost like a "black box" despite producing very deep features. All the architectures we

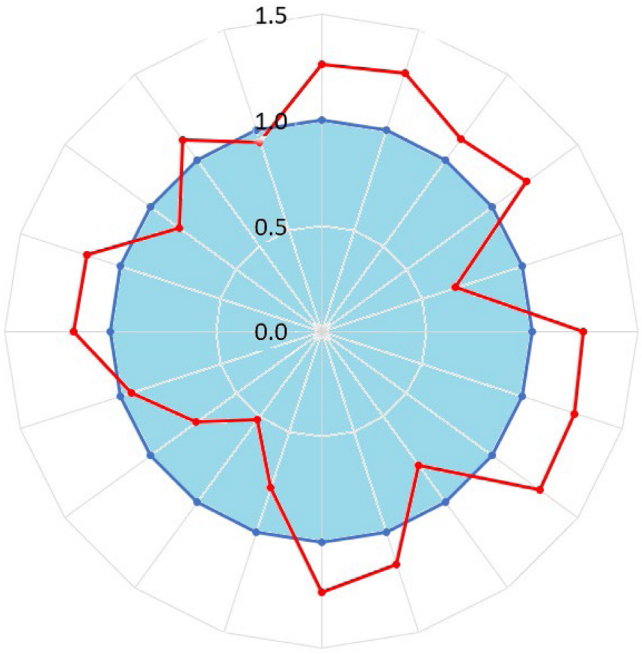


**Fig. 6.** Radar plot for the PvsH-3C sub-classifier.
PvsH: Positive versus Healthy; 3C: Cough vocal task. Radar plot was built on the top 20 features (as ranked by the linear wrapped SVM ranker), averaged over all the subjects, and normalized by the H class. Blue unit circle (colored area) represents the H class, red curve represents the P class. .

**Table 4**
Most relevant feature domains as retained after the wrapper-based ranking step in the Adaboost-based ML pipeline.

| | 1E | 2S | 3C |
|---|---|---|---|
| PvsH comparison | Voicing Probability<br>RASTA-PLP<br>MFCC | RASTA-PLP<br>Spectral Loudness<br>MFCC | RASTA-PLP<br>Spectral Variation<br>Spectral Loudness |
| RvsH comparison | Spectral Variation<br>Energy<br>RASTA-PLP | MFCC<br>RASTA-PLP<br>Energy | RASTA-PLP<br>MFCC<br>Spectral Variation |

Abbreviations: PvsH: Positive versus Healthy; RvsH: Recovered versus Healthy; 1E: Sustained vowel /e/ vocal task sub-classifier; 2S: Sentence vocal task sub-classifier; 3C: Cough vocal task sub-classifier; RASTA-PLP: Features related to RASTA (Relative Spectral) Coefficients applied to the PLP domain (Perceptual Linear Predictive); Spectral Variation: Umbrella term for features related to variations in the spectrum, such as: slope, kurtosis, skewness, flux; MFCC: Mel-frequency Cepstral Coefficients.

adopted, despite being roughly based on state-of-the-art studies, are custom-made in reason of the difficult task of identifying a specific pathology from multiple sources of human voice signals. For the first approach (ML), bagged decision trees (Random Forests) are embedded within a boosting architecture. While bagging aims to cover many permutations of the training set to produce a generalized "tree of trees", boosting the resulting models allows to build the best performing and least biased one. Customized feature sets, resulted from a preliminary correlation-based large-scale skimming followed by a more problem-specific wrapped SVM-based ranker, were used for the training of each Adaboost. The features were only selected on the basis of the training set, meaning that the final validation set was never analyzed by the Adaboost models. On the other hand, the proposed CNN architectures were custom-built, without being based on any specific previous study, and with the added aim to be reasonably "light" for ease of use and to avoid future implementation issues, also considering the absence of improvements when using transfer learning. The substantial acoustic differences between vocalized signals and cough led to the construction of a specific, shallower, CNN for the latter, considering the more straightforward and homogeneous characteristics of those specific recordings, which present less variation within the duration of the audio file and cough section. All of the proposed architectures were then re-applied on the independent validation set.

The results of the Adaboost and CNN-based approaches appear similar for the PvsH comparison, with respective accuracies of 100% and 95%. The RvsH comparison showed Adaboost being significantly more accurate, with 86.1% versus the 75% obtained by the CNN. Interestingly, both approaches for PvsH and RvsH yielded 100% of sensitivity.

The accuracy of the PvsH comparison might prove the feasibility of voice-based COVID-detection, which has solid clinical bases already since most pulmonary diseases have been demonstrated to have distinct and detectable effects on the speech. Moreover, the possibility to also discriminate recovered subjects is in line with the fact that COVID-19 may induce long-lasting damages to the phonatory system, as shown by Helding et al. [76] in a recent study.

By considering the confusion matrices for the single sub-classifiers are concerned, it appeared that even sub-optimal results – like the 66.7% accuracy for the 1E task in the Adaboost-based RvsH comparison – can lead to satisfactory final accuracies when unified with other tasks. This confirms the potential of using more than one vocal indicator. It is also interesting to note how the vowel sound/e/ is the most effective at discriminating positive subjects from healthy ones, whereas it becomes the least promising for the RvsH comparison. This may suggest that the effects of COVID-19 on the voice are subject to change through the course of the disease and recovery. RASTA-PLP processing is assessed as the most recurring domain in the top-ranked features, corroborating and refining the existing approaches mainly based on MFCC, both in the definition of the features to extract for building classifiers and as an alternative to spectrograms for CNN's. Thus, RASTA-PLP processing could be a very viable solution for voice analysis in healthcare and speech recognition, also due to its noise-robust nature [35]. Thus, a more in-depth study of its potential is one of the aims of our future research, especially towards a solid employment of RASTA-temporal diagrams as inputs of a CNN.

Other features like Jitter or HNR, widely used in voice analysis, leave space to more specific features in the frequency domain, here encompassed under the name of "spectral variation", and generally referring to considerations on the slope, kurtosis, prominence of the spectral curves.

In order to consider the features as reliable biomarkers, bias in selection and/or classification should be minimized. We tackled this with the gathering of a polished, well-organized dataset, and with a posterior confrontation of the features with those related to non-pathological effects [77] like age and gender [46] which, if present, could have represented a risk of bias. Despite the hierarchy of accuracies for each sub-classifier being the same between Adaboost and CNN, different subjects get mis-classified in the two approaches. This confirms that, although they probably rely on similar information for the inference, in the end the dynamics of the algorithms are different.

With all these premises, the construction of a tool for remote screening can be reasonably foreseen. In fact, a project is currently ongoing, involving ML-based real-time voice analysis on-site [78]. In these regards, a first concern could be represented by the choice of a suitable environment for voice recordings. The use of adequately quiet rooms for the present research project can easily be replicated in on-site screening facilities, by keeping the room devoid of noisy machinery, and choosing environments

**Table 5**
Literature review.

| Study | Input signals | Recording specifications | COVID-19 screening and validation characteristics | No. of positive (P) subjects | Classes considered | Algorithm(s) | Validation method | Accuracy |
|---|---|---|---|---|---|---|---|---|
| Ours | Vowel, speech, cough | Unique device (lossless) | PCR, serology | 70 (310 total) | P, H, R | Adaboost, CNN (custom) | Independent test set | 100% 86% (R vs H) |
| Laguarta et al. [20] | Cough | Crowdsourced | None (self-reported) | 2660 (5320 total) | P, H | CNN (ResNet50) | Independent test set | 97% |
| Imran et al. [21] | Cough | Unspecified | Unspecified | 70 (543 total) | P, H, pertussis, bronchitis | SVM | Cross-validation | 92% |
| Pinkas et al. [22] | Vowel, speech, cough | Multiple devices (lossy) | PCR | 29 (88 total) | P, H | RNN + SVM | Independent test set | 79% (F1 score) |
| Shimon et al. [23] | Vowel, cough | Multiple devices (lossy) | PCR | 69 (199 total) | P, H | SVM, RF | Independent test set | 80% (mean) |
| Suppakitjanu-sant et al. [25] | Vowel, speech, cough | Unique device | Unspecified | None (76 recovered, 116 total) | R, H | CNN (transfer learning) | Cross-validation | 74% (mean, R vs H) |
| Despotovic et al. [24] | Vowel, speech, breath, cough | Crowdsourced | None (self-reported) | 84 (1103 total) | P, H | Adaboost, Multilayer Perceptron, CNN | Cross-validation | 88% |
| Muguli et al. [26] – DiCOVA challenge | Vowel, speech, breath, cough | Crowdsourced | None (self-reported) | 60 (990 total) | P, H | Various | Various | 73% (baseline) 87% (best) |

Abbreviations: PCR: Polymerase Chain Reaction-based molecular swab; P: COVID-19 Positive subjects; H: Healthy subjects; R: Recovered subjects; CNN: Convolutional Neural Network; SVM: Support Vector Machine; RNN: Recurrent Neural Network.
"Lossless" refers to raw, unprocessed and uncompressed sound data, while "lossy" implies that compression and/or artifacts are present. "Accuracy" refers to the highest reported classification accuracy for the binary Positive VS Healthy classification, except when otherwise specified. Please note that the algorithms used in each study are greatly summarized in the Table. For studies which did not have a single, final, accuracy result, the mean accuracy has been reported, and specified as such.

relatively protected from traffic and crowd noises. Moreover, we also consider quiet domestic rooms to be reasonably close to our experimental settings. The heavy usage of noise-robust acoustic features, like RASTA, also guarantee a certain insensitiveness of the Adaboost-based approach to changes in background noises and/or microphones.

To build a reliable tool, the three classes considered in this study should also be unified in a single classifier. This could be done with a multi-class model or, especially for the ML approach, with the ensemble of three one-vs-one classifiers. However, since the identification of recovered subjects could arguably be less crucial in on-site screening situations, the merging of the H and R classes is also possible, leading to a single binary classification of positives versus non-positives. Both approaches are currently being tested, especially the latter, within the abovementioned national project.

It is worth noting that the choice of recording through smartphones is justified by the need for a widespread and easily accessible strategy for remote screening. Moreover, smartphones have already been proved to be reliable tools for ML-based speech analyses [46,66,79].

The evolution of the disease could itself represent a limitation for the present study, as COVID-19 clinical spectrum is evolving, and pauci-symptomatic positive subjects are becoming more widespread. Besides, long-term effects of the disease on the voice could also be expected to evolve, and a further study of recovered subjects would also be helpful in this matter. Moreover, small datasets constitute a very common problem in bio-engineering ML tasks, and the collection of high-quality data is undoubtedly made difficult by the critical conditions of healthcare institutions. However, we believe that a knowledge-based approach accompanying the data-driven inference can compensate at least some of the typical critical issues of such a task.

### 5.1. Future developments

A more thorough a-priori and posterior analysis of the acoustic features would be beneficial in the future, in order to identify not only the most powerful features for this task, but also a reasonably generalized, unbiased and more definitive set. Furthermore, many refined solutions exist in the neural network realm, especially with regards to pre-trained networks and node splitting, which could lead to better performing architectures, possibly addressing the problem with an even more knowledge-based approach without intensifying too much the calculation burden.

In these regards, the collection of a clean and relatively homogeneous dataset, suitable pre-processing techniques, posterior analysis of features and the usage of two independent state-of-the-art algorithms (specifically chosen and tuned for this task) may hopefully dispel some skepticism towards this pioneering screening technology.

Still, it is important to stress that a screening tool can only offer a preliminary result, suggesting a more extensive validation in case of a positive outcome through conventional diagnostics. This explains our preference for a high sensitivity, despite a certain prevalence for negatively tested subjects in the current diagnostic situation.

All the above-mentioned steps, as already stressed, would be greatly supported by the collection of larger datasets, which is one of our main focuses for the immediate future.

## 6. Conclusions

Our work concerns a novel approach in discriminating three groups of subjects with different COVID-19 status (positive, healthy and recovered), analyzing their vocal performances (sustained vowel, sentence, cough) employing ML algorithms. In order to minimize external biases in the classifiers, we focused on the acquisition of a professionally recorded voice dataset rather than crowdsourced data, which could not only guarantee the maximum reliability of the samples, but also a rigorous annotation and medically proven metadata. On the basis of current state-of-the-art technologies, two algorithms were used following specific fine-tuning and customization. Specifically, the first

approach involves an AdaBoost algorithm with Random Forest weak learners applied to a selection of acoustic features reduced by a CFS followed by an SVM-wrapper-based ranking. The second approach is based on CNN applied to spectrogram images with a knowledge-based data augmentation.

Two binary comparisons, COVID-19 positive versus healthy subjects (PvsH) and recovered versus healthy (RvsH) were considered, with three sub-classifiers per comparison, one for each speech task. Majority voting was used to determine the final results of the comparisons over the three sub-classifiers. Two custom CNN architectures are proposed, one strictly focused on the analysis of the cough sound.

The accuracies are interestingly high, especially for the PvsH comparison, and the Adaboost approach scores higher in both comparisons. Furthermore, a 100% sensitivity for the identification of positive and recovered subjects is obtained by both approaches.

According to our results, traditional ML algorithms proved to still be powerful tools in voice analysis, possibly leading the way for small datasets, over more complex DL solutions. Moreover, we stressed the importance of a knowledge-based approach in such tasks. Carefully built datasets in quiet environments, with adequate pre-processing and fine-tuning of the algorithms, are crucial for a more effective analysis. We observed that the voice sound may hold a COVID-19 "signature", even when the infection is not detectable anymore, which leads to believe that vocal tests can represent a meaningful tool for multiple purposes, including mass-screening, identification of COVID-19 positive subjects and the study of mid and short-term effects of this dreadful disease on the voice.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Authors Giovanni Costantini, Giovanni Saggio, and Antonio Pisani are advisory members of VoiceWise S.r.l., spin-off company of University of Rome Tor Vergata (Rome, Italy) developing voice analysis solutions for diagnostic purposes; Valerio Cesarini cooperates with VoiceWise and is employed by CloudWise S.r.l., a company developing cloud data storage and software solutions.

## Data statement

Clinical data and audio files are not publicly available due to privacy and consent restrictions. Moreover, data contain potentially identifying or sensitive patient information. However, they may be made available to research institutions by the authors upon reasonable request.

## Acknowledgments

The authors would like to thank all volunteers for kindly supporting the present research project as well as the hospitals and staff that made the data collection possible. This work was supported by VoiceWise S.r.l. (spinoff of the University of Rome Tor Vergata) and by the HERMES project funded by the European Space Agency (ESA).

## References

[1] A. Suppa, F. Asci, G. Saggio, P. Di Leo, Z. Zarezadeh, G. Ferrazzano, G. Ruoppolo, A. Berardelli, G. Costantini, Voice analysis with machine learning: one step closer to an objective diagnosis of essential tremor, Mov. Disorders: Off. J. Mov. Disorder Soc. 36 (6) (2021) 1401–1410, http://dx.doi.org/10.1002/mds.28508.

[2] J.P. Teixeira, J. Fernandes, F. Teixeira, P.O. Fernandes, Acoustic analysis of chronic laryngitis-statistical analysis of sustained speech parameters, in: 11th International Joint Conference on Biomedical Engineering Systems and Technologies, 2018, pp. 168–175.

[3] S.C. Costa, B.G.A. Neto, J.M. Fechine, M. Muppa, Short-term cepstral analysis applied to vocal fold edema detection, in: BIOSIGNALS (2), 2008, pp. 110–115.

[4] M. Petrovic-Lazic, N. Jovanovic, M. Kulic, S. Babac, V. Jurisic, Acoustic and perceptual characteristics of the voice in patients with vocal polyps after surgery and voice therapy, J. Voice 29 (2) (2015) 241–246.

[5] M. Alves, E. Krüger, B. Pillay, K. Van Lierde, J. Van der Linde, The effect of hydration on voice quality in adults: A systematic review, J. Voice 33 (1) (2019) 125–e13.

[6] T. Zacharia, S. Rao, S.K. Hegde, P. D'souza, J. James, M.S. Baliga, Evaluation of voice parameters in people with head and neck cancers: an investigational study, Middle East J. Cancer 7 (4) (2016) 193–197.

[7] M. Alhussein, Monitoring Parkinson's disease in smart cities, IEEE Access 5 (2017) 19835–19841.

[8] P. Gómez-Vilda, A.R.M. Londral, V. Rodellar-Biarge, J.M. Ferrández-Vicente, M. de Carvalho, Monitoring amyotrophic lateral sclerosis by biomechanical modeling of speech production, Neurocomputing 151 (2015) 130–138.

[9] G. Albertini, S. Bonassi, V. Dall'Armi, I. Giachetti, S. Giaquinto, M. Mignano, Spectral analysis of the voice in Down syndrome, Res. Dev. Disabil. 31 (5) (2010) 995–1001.

[10] V. Pareek, R.K. Sharma, Coronary heart disease detection from voice analysis, in: 2016 IEEE Students' Conference on Electrical, Electronics and Computer Science, SCEECS, IEEE, 2016, pp. 1–6.

[11] J.E. Oh, Y.M. Choi, S.J. Kim, C.U. Joo, Acoustic variations associated with congenital heart disease, Korean J. Pediatr. 53 (2) (2010) 190–194.

[12] M. Sakai, Feasibility study on blood pressure estimations from voice spectrum analysis, Int. J. Comput. Appl. 109 (7) (2015) 39–43.

[13] F. Asci, G. Costantini, G. Saggio, A. Suppa, Fostering voice objective analysis in patients with movement disorders, Mov. Disorders 36 (4) (2021) 1041.

[14] M. Todisco, E. Alfonsi, S. Arceri, G. Bertino, C. Robotti, M. Albergati, M. Gastaldi, C. Tassorelli, G. Cosentino, Isolated bulbar palsy after SARS-CoV-2 infection, Lancet. Neurol. 20 (3) (2021) 169–170, http://dx.doi.org/10.1016/S1474-4422(21)00025-9.

[15] G. Hu, X. Peng, Y. Yang, T.M. Hospedales, J. Verbeek, Frankenstein: learning deep face representations using small data, IEEE Trans. Image Process. 27 (1) (2018) 293–303, http://dx.doi.org/10.1109/TIP.2017.2756450.

[16] D. Sztahó, K. Gábor, T. Gábriel, Deep learning solution for pathological voice detection using LSTM-based autoencoder hybrid with multi-task learning, in: Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 2: BIOSIGNALS, ISBN: 978-989-758-490-9, 2021, pp. 135–141, http://dx.doi.org/10.5220/0010193101350141.

[17] I. Nissar, W.A. Mir, Izharuddin, T.A. Shaikh, Machine learning approaches for detection and diagnosis of parkinson's disease - a review, in: 2021 IEEE, 7th International Conference on Advanced Computing and Communication Systems, ICACCS, 2021, 2021, pp. 898–905, http://dx.doi.org/10.1109/ICACCS51430.2021.9441885.

[18] A. Benba, A. Jilbab, A. Hammouch, S. Sandabad, Voiceprints analysis using MFCC and SVM for detecting patients with Parkinson's disease, in: 2015 International Conference on Electrical and Information Technologies, ICEIT, 2015, pp. 300–304.

[19] A. Suppa, F. Asci, G. Saggio, L. Marsili, D. Casali, Z. Zarezadeh, G. Ruoppolo, A. Berardelli, G. Costantini, Voice analysis in adductor spasmodic dysphonia: Objective diagnosis and response to botulinum toxin, Parkinsonism Rel. Disord. 73 (2020) 23–30, http://dx.doi.org/10.1016/j.parkreldis.2020.03.012.

[20] J. Laguarta, F. Hueto, B. Subirana, COVID-19 artificial intelligence diagnosis using only cough recordings, IEEE Open J. Eng. Med. Biol. (2020).

[21] A. Imran, I. Posokhova, H.N. Qureshi, U. Masood, S. Riaz, K. Ali, M. Nabeel, AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app, 2020, arXiv preprint arXiv:2004.01275.

[22] G. Pinkas, Y. Karny, A. Malachi, G. Barkai, G. Bachar, V. Aharonson, SARS-CoV-2 detection from voice, IEEE Open J. Eng. Med. 1 (2020) 268–274, 2020.

[23] C. Shimon, G. Shafat, I. Dangoor, A. Ben-Shitrit, Artificial intelligence enabled preliminary diagnosis for COVID-19 from voice cues and questionnaires, J. Acoust. Soc. Am. 149 (2) (2021) 1120.

[24] V. Despotovic, M. Ismael, M. Cornil, R.M. Call, G. Fagherazzi, Detection of COVID-19 from voice, cough and breathing patterns: Dataset and preliminary results, Comput. Biol. Med. 138 (2021) 104944.

[25] P. Suppakitjanusant, S. Sungkanuparph, T. Wongsinin, et al., Identifying individuals with recent COVID-19 through voice classification using deep learning, Sci. Rep. 11 (2021) 19149, http://dx.doi.org/10.1038/s41598-021-98742-x.

[26] A. Muguli, L. Pinto, R. Nirmala, N. Sharma, P. Krishnan, P.K. Ghosh, R. Kumar, S. Bhat, S.R. Chetupalli, S. Ganapathy, S. Ramoji, N. Viral, DiCOVA Challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics, 2021, arXiv:2103.09148, Pre-print within the Interspeech 2021 challenge.
[27] C. Robotti, G. Costantini, G. Saggio, V. Cesarini, A. Calastri, E. Maiorano, D. Piloni, T. Perrone, U. Sabatini, V.V. Ferretti, I. Cassaniti, F. Baldanti, A. Gravina, A. Sakib, E. Alessi, M. Pascucci, D. Casali, Z. Zarezadeh, V. Del Zoppo, A. Pisani, M. Benazzo, Machine learning-based voice assesment for the detection of positive and recovered COVID-19 patients, J. Voice (2021) http://dx.doi.org/10.1016/j.jvoice.2021.11.004, in press.
[28] Fant G., Acoustic Theory of Speech Production, Mouton, The Hague, 1960.
[29] MATLAB, 9.7.0.1190202 (R2019b), The MathWorks Inc., Natick, 2018.
[30] S. Glen, Lowess smoothing in statistics: what is it? From: StatisticsHowTo.com: elementary statistics for the rest of us! 2013, https://www.statisticshowto.com/lowess-smoothing/.
[31] F. Eibe, M. Hall, I. Witten, The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques, Fourth ed., Morgan Kaufmann, 2016.
[32] F. Eyben, M. Wöllmer, M Björn Schuller, OpenSMILE - the munich versatile and fast open-source audio feature extractor, in: Proc. ACM Multimedia (MM), ACM, Florence, Italy, ISBN: 978-1-60558-933-6, 2010, pp. 1459–1462, 25.-29.10.2010.
[33] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini, The INTERSPEECH 2016 computational paralinguistics challenge: deception, sincerity and native language, 2016.
[34] B.P. Bogert, M.J.R. Healy, J.W. Tukey, The quefrency alanysis [sic] of time series for echoes: cepstrum, pseudo autocovariance, cross-cepstrum and saphe cracking, in: M. Rosenblatt (Ed.), Proceedings of the Symposium on Time Series Analysis, 1963.
[35] H. Hermansky, N. Morgan, RASTA processing of speech, IEEE Trans. Speech Audio Process. 2 (1994) 578–589, http://dx.doi.org/10.1109/89.32661.
[36] S. Yeldener, Method of determining the voicing probability of speech signals - united states patent USOO637792OB2, Patent No.: US 6, 377, 920 B2, Apr. 23, 2002.
[37] A.K. Anweiler, J.L. Verhey, Spectral loudness summation for short and long signals as a function of level, J. Acoust. Soc. Am. 119 (5 Pt 1) (2006) 2919–2928, http://dx.doi.org/10.1121/1.2184224.
[38] M. Köppen, The curse of dimensionality, 2009, http://dx.doi.org/10.1007/978-0-387-39940-9_133.
[39] A. Salimi, M. Ziaii, A. Amiri, M.H. Zadeh, S. Karimpouli, M. Moradkhani, Using a feature subset selection method and support vector machine to address curse of dimensionality and redundancy in Hyperion hyperspectral data classification, Egypt. J. Remote Sens. Space Sci. 21 (2018) 27–36, http://dx.doi.org/10.1016/j.ejrs.2017.02.003.
[40] M.A. Hall, Correlation-Based Feature Selection for Machine Learning, University of Waikato, Department of Computer Science, Hamilton, NZ, 1999.
[41] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, Introduction to Algorithms, MIT Press, Cambridge, 2009.
[42] C. Cortes, V.N. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[43] J. Platt, Sequential minimal optimization: a fast algorithm for training support vector machines, 1998.
[44] I. Hammami, L. Salhi, S. Labidi, Pathological voices detection using support vector machine, 2016, pp. 662–666.
[45] J.I. Godino-Llorente, P. Gómez-Vilda, N. Sáenz-Lechón, M. Blanco-Velasco, F. Cruz-Roldán, M.A. Ferrer-Ballester, Support vector machines applied to the detection of voice disorders, in: Faundez-Zanuy M. Janer L. Esposito A. Satue-Villar A. Roure J. Espinosa-Duro V. (Ed.), Nonlinear Analyses and Algorithms for Speech Processing, NOLISP 2005, in: Lecture Notes in Computer Science, vol. 3817, 2006.
[46] F. Asci, G. Costantini, P. Di Leo, A. Zampogna, G. Ruoppolo, A. Berardelli, G. Saggio, A. Suppa, Machine-learning analysis of voice samples recorded through smartphones: the combined effect of ageing and gender, Sensors (Basel, Switzerland) 20 (18) (2020) 5022, http://dx.doi.org/10.3390/s20185022.
[47] X. Zhang, L. Zhang, Z. Tao, H. Zhao, Acoustic characteristics of normal and pathological voices analysis and recognition, in: 2019 6th International Conference on Systems and Informatics, ICSAI, 2019, pp. 1423–1427.
[48] M.F. Delgado, E. Cernadas, S. Barro, D.G. Amorim, Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. 15 (2014) 3133–3181.
[49] R. Sheibani, E. Nikookar, S.E. Alavi, An ensemble method for diagnosis of parkinson's disease based on voice measurements, J. Med. Signals Sens. 9 (4) (2019) 221–226, http://dx.doi.org/10.4103/jmss.JMSS_57_18, Published 2019 Oct 24.
[50] A.J. Wyner, M.L. Olson, J. Bleich, D. Mease, Explaining the success of AdaBoost and random forests as interpolating classifiers, 2017, ArXiv, abs/1504.07676.
[51] J. Thongkam, G. Xu, Y. Zhang, AdaBoost algorithm with random forests for predicting breast cancer survivability, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008, pp. 3062–3069.
[52] Y. Freund, R.E. Schapire, Experiments with a new boosting algorithm, 1996.
[53] H. Parmar, S. Bhanderi, G. Shah, Sentiment mining of movie reviews using random forest with tuned hyperparameters, in: Conference: International Conference on Information Science, Kerala, 2014.
[54] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, http://dx.doi.org/10.1023/A:1010933404324.
[55] G. Leshem, Y. Ritov, Traffic flow prediction using adaboost algorithm with random forests as a weak learner, J. Int. J. Intell. Technol. 2 (2007) 1305–6417, 2007.
[56] G. Gosztolya, R. Busa-Fekete, T. Grósz, L. Tóth, DNN-based feature extraction and classifier combination for child-directed speech, cold and snoring identification, in: INTERSPEECH, 2017.
[57] V. Bansal, G. Pahwa, N. Kannan, Cough Classification for COVID-19 based on audio mfcc features using Convolutional Neural Networks, in: 2020 IEEE International Conference on Computing, Power and Communication Technologies, GUCON, 2020, pp. 604–608, http://dx.doi.org/10.1109/GUCON48875.2020.9231094.
[58] N. Cummins, A. Baird, B.W. Schuller, Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning, Methods (ISSN: 1046-2023) 151 (2018) (2018) 41–54, http://dx.doi.org/10.1016/j.ymeth.2018.07.007.
[59] H. Nguyen, C. Nguyen, T. Ino, B. Indurkhya, M. Nakagawa, Text-independent writer identification using convolutional neural network, Pattern Recognit. Lett. 121 (2018) http://dx.doi.org/10.1016/j.patrec.2018.07.022.
[60] B.B. Monson, E.J. Hunter, A.J. Lotto, B.H. Story, The perceptual significance of high-frequency energy in the human voice, Front. Psychol. 5 (2014) 587, http://dx.doi.org/10.3389/fpsyg.2014.00587.
[61] G. Marcus, Deep learning: a critical appraisal, 2018, ArXiv, abs/1801.00631.
[62] L. Nanni, G. Maguolo, M. Paci, Data augmentation approaches for improving animal audio classification, 2020, ArXiv, abs/1912.07756.
[63] A.H. Barshooi, A. Amirkhani, A novel data augmentation based on gabor filter and convolutional deep learning for improving the classification of COVID-19 chest X-ray images, Biomed. Signal Process. Control 72 (2022) 103326, http://dx.doi.org/10.1016/j.bspc.2021.103326.
[64] J. Chenou, G. Hsieh, Increasing the robustness of deep learning with colored noise augmentation, 2019.
[65] D.S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E.D. Cubuk, Q.V. Le, SpecAugment: a simple data augmentation method for automatic speech recognition, in: INTERSPEECH, 2019.
[66] M. Pahar, M. Klopper, R. Warren, T. Niesler, COVID-19 cough classification using machine learning and global smartphone recordings, 2, 2020, arXiv:2012.01926v1 [cs.SD].
[67] D. Kingma, J. Ba, Adam: a method for stochastic optimization, in: International Conference on Learning Representations, 2014.
[68] P. Bühlmann, S Van De Geer, Statistics for High-Dimensional Data, in: Springer Series in Statistics, vol. 9, ISBN: 978-3-642-20191-2, 2011, http://dx.doi.org/10.1007/978-3-642-20192-9.
[69] J. Fernandes, L. Silva, F. Teixeira, V. Guedes, J. Santos, J. Teixeira, Parameters for vocal acoustic analysis - cured database, Procedia Comput. Sci. 164 (2019) 654–661, http://dx.doi.org/10.1016/j.procs.2019.12.232.
[70] K.D. Bartl-Pokorny, F.B. Pokorny, A. Batliner, S. Amiriparian, A. Semertzidou, F. Eyben, E. Kramer, F. Schmidt, R. Schönweiler, M. Wehler, B.W. Schuller, The voice of COVID-19: Acoustic correlates of infection in sustained vowels, J. Acoust. Soc. Am. 149 (6) (2021) 4377, http://dx.doi.org/10.1121/10.0005194.
[71] M. Tamazin, A. Gouda, M. Khedr, Enhanced automatic speech recognition system based on enhancing power-normalized cepstral coefficients, Appl. Sci. 9 (2019) http://dx.doi.org/10.3390/app9102166.
[72] R. Singh, P. Rao, Spectral subtraction speech enhancement with RASTA filtering, in: Proceding of National Conference on Communications, NCC, Kanpur, India, 2007, 2007.
[73] V. Cesarini, N. Casiddu, C. Porfirione, G. Massazza, G. Saggio, G. Costantini, A machine learning-based voice analysis for the detection of dysphagia biomarkers, in: 2021 IEEE International Workshop on Metrology for Industry 4.0 IoT (MetroInd4.0 IoT), 2021, pp. 407–411, http://dx.doi.org/10.1109/MetroInd4.0IoT51437.2021.9488503.
[74] R.D. Kent, R.L. Sufit, J.C. Rosenbeck, J.F. Kent, G. Weismer, R.E. Martin, B.R. Brooks, Speech deterioration in amyotrophic lateral sclerosis: a case study, J. Speech Hear. Res. 34 (6) (1991) 1269–1275, http://dx.doi.org/10.1044/jshr.3406.1269.
[75] G. Saggio, Are sensors and data processing paving the way to completely non-invasive and not-painful medical tests for widespread screening and diagnosis purposes? in: BIODEVICES, 2020, pp. 207–214.
[76] L. Helding, T.L. Carroll, J. Nix, M.M. Johns, W.D. LeBorgne, D. Meyer, COVID-19 after effects: concerns for singers, J. Voice:

Off. J. Voice Found. (2020) http://dx.doi.org/10.1016/j.jvoice.2020.07.032, S0892-1997(20)30281-2. Advance online publication.
[77] G. Saggio, G. Costantini, Worldwide healthy adult voice baseline parameters: a comprehensive review, J. Voice (2020).
[78] HERMES Project - https://www.leonardocompany.com/en/news-and-stories-detail/-/detail/hermes-the-telespazio-and-e-geos-solution-responding-to-healthcare-needs.
[79] V. Uloza, E. Padervinskis, A. Vegiene, R. Pribuisiene, V. Saferis, E. Vaiciukynas, A. Gelzinis, A. Verikas, Exploring the feasibility of smart phone microphone for measurement of acoustic voice parameters and voice pathology screening, in: European Archives of Oto-Rhino-Laryngology: Official Journal of the European Federation of Oto-Rhino-Laryngological Societies, EUFOS, 2015.