Biomedical Signal Processing and Control 104 (2025) 107530

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Towards an explainable Artificial intelligence system for voice pathology identification and post-treatment characterisation

Federico Calà [a,*], Lorenzo Frassineti [a], Giovanna Cantarella [b,c], Giulia Buccichini [c], Ludovica Battilocchi [b], Claudia Manfredi [a], Antonio Lanatà [a]

[a] Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy
[b] IRCCS Ca' Grande Foundation, Ospedale Maggiore Policlinico Milano, Milano, Italy
[c] Department of Clinical Sciences and Community Health, Università degli Studi di Milano, Milano, Italy

ARTICLE INFO

Keywords:
Artificial Intelligence
Machine Learning
Acoustic Analysis
BioVoice
Interpretable AI
Dysphonia
Benign lesions
Unilateral Vocal Fold Paralysis Post-treatment

ABSTRACT

The voice pathology identification task has recently gained great attention. However, several research questions remain open. This study proposes an explainable AI framework to address the implicit role of age in voice pathology recognition and to investigate vocal quality improvement after surgical treatment in organic voice disorders. The aim is also to define an optimal features subset through predictor importance analysis. A set of 287 patients diagnosed with benign lesions of vocal folds (BLVF) and unilateral vocal fold paralysis (UVFP) was enrolled. Classification experiments were performed for female (F) and male (M) groups: they aimed at distinguishing BLVF from UVFP in age-unbalanced (E1) and age-balanced (E2) datasets, differentiating BLVF subclasses (E3), and detecting pre- and post-treatment conditions (E4). The comparison between E1 and E2 suggests that age does not influence the classification performance. In E1, 76% (F) and 81% (M) accuracies were obtained. The best features concerned vocal fold dynamics and articulator positioning for F and M datasets. In E3, an accuracy of 60% was achieved, suggesting that larger datasets are required. In E4, the best models showed 76% (F) and 72% (M) accuracy, with a good sensitivity in detecting pre-treatment patients. The error rate analysis proved that UVFP was the most misclassified group. Moreover, an agreement between the AI outcome and perceptual evaluations was detected for misclassified recordings. These results suggest their clinical relevance to highlight key aspects of voice quality recovery and to define acoustic parameters that otolaryngologists could employ to monitor the patient's follow-up.

1. Introduction

Dysphonia is a medical condition characterised by higher irregular pitch, lower vocal quality, and loudness concerning gender- and age-matched normophonic subjects [1], representing one of the most evident markers of voice disorders. Its prevalence in the adult population is estimated between 3 % and 10 % [2–4], but many demographic factors contribute to the rise of this value. Dysphonia prevalence in teachers can indeed increase up to 25 % [5], and depends on gender, with a female/male ratio of 1.5:1 [3]. Moreover, prevalence depends on etiopathogenesis, which includes tissue infections and surface irritation, mechanical stress (which may provoke the formation of vocal nodules and polyps), tissue changes (for instance, cysts), neuromuscular anomalies (such as vocal fold paralysis, spasmodic dysphonia or Parkinson's disease), cognitive impairment (e.g., Alzheimer's disease), and psychogenic causes (like chronic stress disorders or bipolarity).

The clinical diagnosis of dysphonia is based on the direct inspection of the vocal folds by laryngoscopy, which represents the gold standard method. Here, complementary assessments are proposed to consider the lack of high- resolution endoscopy in decentralised ambulatories or primary care units. Moreover, laryngoscopy requires being physically present in hospitals, which can be demanding for fragile subjects [6] and in case of worldwide health emergencies such as the COVID-19 pandemic [7].

Several studies reported how voice disorders can significantly hinder an individual's quality of life due to impaired communication and lead to professional and social problems, anxiety, isolation, and depression [8,9]. Nevertheless, subjects that seek treatment are pretty few, determining a relevant pathology underdiagnosis [4,10].

Several indirect methods are available to assess voice disorders to

* Corresponding author.
E-mail address: federico.cala@unifi.it (F. Calà).

https://doi.org/10.1016/j.bspc.2025.107530
Received 20 September 2024; Received in revised form 28 December 2024; Accepted 17 January 2025
Available online 26 January 2025

reduce diagnosis invasiveness and improve tolera- bility [11,12]. Perceptual voice evaluation is a well-established method typically performed along with laryngoscopy.

Nevertheless, it suffers from inter-rater variability and physicians' experience limitations [13].

Another objective method to evaluate voice features combines acoustic analysis based on the parametrisation of voice and speech production through linear and nonlinear signal processing methods and artificial intelligence (AI) techniques. This combined approach would allow the development of reliable tools to support diagnosis, monitor vocal properties after surgical or logopedic treatments, and detect early symptoms of voice pathology. As pointed out by Godino-Llorente et al. [1], one of the first aspects to consider for the development of automatic systems of voice analysis is the choice of the input utterance, which indeed plays an essential role in feature extraction, and it is strongly related to the type of pathologies of interest. Most studies rely on the analysis of the sustained vowel /a/ as its phonation requires a relaxed and open vocal tract and stable tongue and jaw position [14]. It also demonstrated relative independence from intonation, dialects, and language [1]. However, using a single vowel might not fully reflect voice characteristics. Furthermore, it does not account for some aspects of speech production that other tasks may reveal [6]. Therefore, other tasks have been proposed, such as:

- Running speech. It contains important information about voice breaks, onset, and offset. It can be crucial for pathology detection and identification. Ali et al. [6] extracted linear prediction coefficients (LPC) from spontaneous speech and achieved high performances with a Gaussian Mixture Model in distinguishing healthy and disordered speech. Godino-Llorente et al. [15] have obtained similar results using features extracted from the standardised "Rainbow Passage" and implementing a multilayer perceptron.
- Number listing task for the evaluation of coarticulation. Frassineti et al. [16] demonstrated that acoustic parameters and entropy indexes can discriminate vocal phenotypes of patients diagnosed with genetic syndromes. In contrast, Ali et al. [17] computed voice contour measures and successfully used them as features in a support vector machine (SVM) to separate pathological and healthy subjects.
- Diadochokinetic task. It assesses the ability to perform the cycles of articulatory movements necessary to form syllables consisting of a consonant and a vowel [18]. Usually, it consists of several consecutive repetitions of the /pa/, /ta/, /ka/ syllables. This task was administered to detect and evaluate the severity degree of Parkinson's disease [19,20] and motor speech disorders [21].

From these tasks, voice pathology classification systems extract a set of parameters that can be conventional acoustic features (e.g., fundamental frequency, jitter, shimmer, harmonic-to-noise ratio), mel-frequency cepstrum coefficients (MFCC) and other parameters (as linear predictive coding coefficients or high-order statistics) to distinguish vocal properties of healthy and pathological subjects [10,22–24]. In this binary problem, performance has achieved high accuracy (up to 95 % [24]). Moreover, the increasing availability of large databases has led to a wide application of deep learning approaches to classify motor imagery from EEG patterns [25,26], to detect cardiovascular diseases [27] and also voice pathologies [28], as well as depression [29] and Parkinson's disease [30] from acoustic recordings. However, properly tuning these models to obtain reliable performance remains challenging as, especially in the healthcare domains, acquiring an adequate sample size may not be guaranteed due to recording difficulties and the rarity of specific pathologies [31]. This is one of the reasons machine learning still represents a robust and feasible methodology for data analysis.

AI-based methods can also be used to develop automatic tools to support the differential diagnosis of voice disorders. Relevant advances in the voice disorder identification task were achieved at the FEMH 2018 challenge and the 2019 IEEE BigDataCup. Degila et al. [32] used MFCC and Support Vector Machine (SVM) to discriminate nodules, polyps, and cysts (collectively referred to as a single class), unilateral vocal fold paralysis (UVFP), and glottis neoplasm on a small cohort of patients, obtaining 97 % sensitivity. These promising results led to more detailed research. Miliaresi and Pikrakis [33] have used these same limited data to develop a four-class deep learning model that achieved 64 % accuracy. Harar et al. [34] extracted conventional acoustic features, spectrogram parameters, and MFCC from the sustained /a/ of patients affected by several voice disorders contained in the Saarbrücken Voice Database [35] and obtained an F1-score of 73 % in separating them into different classes, with an XGBoost Tree. Finally, Marchese et al. [36] performed an exploratory analysis by implementing an SVM classifier to separate nodules, polyps, cysts, and Reinke's oedema that obtained 55 % accuracy for female patients.

Consequently, research is moving toward making black-box models' output more understandable since experts need more information to make a decision, and a simple binary prediction is insufficient to be accepted as clinically relevant. Most of these papers mainly concerned methodological approaches. Indeed, they focused on feature extraction and selection methods and designing machine or deep learning frameworks to increase performance. Notable results have been achieved to the detriment of explainability, which has rarely been addressed. This represents a critical issue since a technique which is not directly interpretable and trustworthy can pose a major threat to the usability of AI-based methods [37]. It is especially true in the healthcare domain, where decision support systems are based on data that are often limited, imbalanced, heterogeneous and noisy [38]. Therefore, to be considered practical tools and to avoid limiting their effectiveness, explainable AI (XAI) has been proposed to allow field experts to understand and manage AI model development and results. Two different types of XAI approaches can be defined [37,39]:

- Post-hoc techniques that provide text, visual or local explanations to allow model interpretability. In voice analysis, permutation feature importance and Shapley values were used to understand which acoustic features were considered relevant by machine learning models to separate patients diagnosed with polyps and UVFP [40]. Shapley values were also computed to find acoustic parameters able to detect Parkinson's disease in patients' standardised speech [41]. Furthermore, the Local-Interpretable Model-agnostic Explanation method was implemented to evaluate the relationships between perceptual and objective assessment of spasmodic dysphonia [42]. These techniques were originally proposed for machine learning models, and research is now starting to modify and implement them also in deep learning-based approaches [43–45]. However, more work is needed to validate such methodologies, and their lack of usage could relevantly hinder this process [43,45].
- AI frameworks that are intrinsically interpretable. Classifiers such as K-Nearest Neighbors or Decision Trees are commonly used for voice pathology detection [1]. Still, their explainability is overshadowed by a usually lower classification performance with respect to more complex but less transparent models such as Support Vector Machine or Ensemble Trees classifiers [39].

Moreover, even when interpretability is addressed, and considering the intended use of computer-aided diagnostic systems, explainability relies also on extracted features, which should be easily related to physiological processes. Specifically, it would be important for otolaryngologists to understand which acoustic features were considered most relevant for a classifier for the distinction of two or more classes and that such metrics could have a reliable physiological meaning (e.g., the first formant is related to the degree of the pharynx constriction [46]). Such an approach would make it easier for clinicians to plan intervention and rehabilitation programs.

Attention to interpretability could also be clinically relevant for assessing voice quality after surgical, pharmacological, or logopedic

treatments. Suppa et al. [47] investigated with an Artificial Neural Network the change in voice quality before and after the injection of botulinum toxin in patients diagnosed with spasmodic dysphonia. Bensoussan et al. [48] used a previously trained classifier on *cis*-male and *cis*-female voices to investigate the efficiency of gender reassignment therapies on a cohort of *trans*-female patients. These studies demonstrated that AI could also be successfully deployed for such purposes. Still, they need further examination to determine which type of therapy can restore or achieve proper voice quality or to analyse how vocal properties change over time.

This study aims to extend previous work [49] on combining acoustic analysis and machine learning techniques to develop an automatic tool based on physiologically explainable vocal characteristics. The study focused on patients diagnosed with nodules, polyps, and cysts (considered separately and jointly as the class of benign lesions of the vocal folds, BLVF) and unilateral vocal fold paralysis (UVFP). These two medical conditions were chosen to test the feasibility of our approach, as they present very different laryngeal dynamics. Moreover, whether classifiers could distinguish pathological voices before and after surgical treatment is investigated. Feature extraction is performed on the sustained vowel /a/ to obtain results comparable with the literature and on an articulation task as a new approach to voice and speech analysis in pathology identification. Similarly to [50], it explored if the joint use of acoustic features extracted from both sustained /a/ and the Italian word /aiuole/ (IPA transcription: «a'jwɔle», English translation: «flowerbeds») can lead to improvements in classification performance. Four experiments were performed separately for both genders to identify BLVF and UVFP in age-unbalanced and age-balanced models, to recognise BLVF subcategories (i.e., nodules, polyps and cysts) and to characterise pre- and post-treatment voice quality. Moreover, for each experiment, a thorough model interpretation is performed by identifying which acoustic parameters were considered by models for class separation and understanding whether these metrics present significant differences between pathologies employing statistical analysis.

The major contributions of this work follow:

- To evaluate the implicit role of age in the voice identification problem. Indeed, it is well-known that ageing affects vocal quality. However, models performing this task have never addressed its potential bias.
- To investigate the capabilities of ML in the BLVF differential diagnosis task, which still constitutes a poorly studied research question.
- To explore with ML methods the improvement of vocal quality after surgical treatment. This is the first time such an approach has been applied to BLVF and UVFP patients.

These results may be used to develop an automatic tool for voice pathology identification that could be effective, cheap, and contactless. This tool may also support otolaryngologists in differential diagnosis and treatment monitoring.

2. Materials and methods

2.1. Database

This study recruited 287 patients (of which 183 were female F, mean age = 44.6 ± 4.6 years, 104 were male M, mean age = 42.6 ± 9.4 years) at the Ospedale Maggiore Policlinico Milano (Milan, Italy). All patients underwent both perceptual evaluations of their voice and videolaryngostroboscopic assessment performed by the same clinician and in the same hospital. Patients were diagnosed as nodules (15F, 2 M), polyps (56F, 38 M), cysts (34F, 6 M), and unilateral vocal fold paralysis (78F, 56 M).

Voice samples were recorded using a C1000S dynamic microphone (AKG, Vienna, Austria) with a sampling frequency of 44.1 kHz and at a fixed distance of 5 cm from the patient's mouth during the phonation, at comfortable pitch and loudness of the sustained cardinal vowel /a/ and the Italian word /aiuole/, which contains all the basic vowels of the Italian language (IPA transcription: «a'jwɔle», English translation: «flowerbeds»). All data was recorded anonymously, and informed consent was obtained from each patient.

Commonly, UVFP patients present a higher mean age than BLVF ones [51]. In our database, female patients diagnosed with UVFP have a mean age of 44 years (minimum–maximum range: 19–68), whereas the ones diagnosed with UVFP present a mean age of 50 years (min–max range: 21–72). For male patients, the mean age for BLVF is 42 years (min–max range: 19–78), whereas the mean age of UVFP subjects is 51 years (min–max range: 26–80). After checking for normality of data distribution, we performed either a *t*-test or Mann-Whitney test with a α-level of significance at 0.05 to understand whether the age difference between BLVF and UVFP cohorts is statistically significant. Firstly, the test was carried out considering all available data. Then, in case of a significant difference, an age subrange between 18 and 60 years (according to [34]) was selected to get a homogeneous group. The distribution of patients, for both age inhomogeneous and homogeneous groups, is displayed in Table 1.

Finally, a dataset of 90 patients' voices (61F, 29 M) that underwent medical intervention was available for analysis. These subjects were a part of the 287 patients database. Audio files were recorded following the same procedure. A perceptual evaluation with the GRB scale [52] was performed blindly by a speech therapist with long experience in voice disorders. The GRB scale is composed of three items:

- G (global grade of dysphonia): the judgement is based on the overall impression of voice quality deterioration.
- R (Roughness): the impression of irregular F0 and noise.
- B (Breathiness): turbulent noise related to air escape through the vocal folds.

The scores ranged from 0 to 3, where 0 is the rating for the best voice condition and 3 for the worst one. For some post- treatment subjects, more than one audio file was recorded after surgery for voice quality follow-up. Table 2 illustrates the overall number of recordings divided by gender and vocal task. The number of audio samples for the BLVF group is reported in brackets.

2.2. Feature extraction

After manual segmentation, audio files were processed using the open-source BioVoice software [53]. This tool provides a user-friendly interface that allows selecting the proper frequency range for adults (distinguishing between male and female as well), children, and newborns to perform acoustic analysis. Specifically, it computes several acoustic parameters in both the time and frequency domain. Table 3 shows those used in this work and their description. Two parameters, namely T0 (F0 min) and T0 (F0 max), were further processed to make them independent from the total duration of the audio recording. Therefore, they were normalised with respect to the time duration of the recording. It allowed obtaining a reliable estimation of the time instant at which the minimum or maximum of the F0 occurs, i.e., at the

Table 1
Patients distribution divided by age, gender and voice pathology. The acronym y.o. stands for years old.

	Female patients		Male patients	
Pathology	Whole database	18–60 y.o. age group	Whole database	18–60 y.o. age group
Nodules	15	7	2	0
Polyps	56	36	38	20
Cysts	34	18	6	4
UVFP	78	43	58	18

Table 2
Number of post-treatment audio recordings for female and male patients divided by vocal task type. In brackets, the number of BLVF recordings is reported.

Gender	/a/	/aiuole/
F	94 (33)	89 (30)
M	61 (12)	57 (9)

Table 3
Acoustic parameters estimated by BioVoice.

Feature	Description
F0 mean [Hz]	Mean fundamental frequency
F0 median [Hz]	Median fundamental frequency
F0 std [Hz]	Standard deviation of fundamental frequency
F0 min [Hz]	Minimum value of the fundamental frequency
T0 (F0 min) [s]	Time instant at which the minimum of F0 occurs
F0 max [Hz]	Maximum value of the fundamental frequency
T0 (F0 max) [s]	Time instant at which the maximum of F0 occurs
Jitter [%]	Irregularity of F0 in the frequency domain
NNE [dB]	Normalized Noise Energy
F1 mean [Hz]	Mean value of the first formant
F1 median [Hz]	Median value of the first formant
F1 std [Hz]	Standard deviation of the first formant
F1 min [Hz]	Minimum value of the first formant
F1 max [Hz]	Maximum value of the first formant
F2 mean [Hz]	Mean value of the second formant
F2 median [Hz]	Median value of the second formant
F2 std [Hz]	Standard deviation of the second formant
F2 min [Hz]	Minimum value of the second formant
F2 max [Hz]	Maximum value of the second formant
F3 mean [Hz]	Mean value of the third formant
F3 median [Hz]	Median value of the third formant
F3 std [Hz]	Standard deviation of the third formant
F3 min [Hz]	Minimum value of the third formant
F3 max [Hz]	Maximum value of the third formant
Signal duration [s]	Time duration of the audio file
% voiced	Percentage of voiced parts inside the whole signal
Voiced duration [s]	Total duration of voiced parts

beginning, in the middle, or at the end of each recording.

These features were chosen due to their direct relationship with physiological processes. Additional parameters such as the ppq5 jitter or the shimmer were not considered as they would need to be calculated with other software, e.g., PRAAT [54]. These systems model voice production differently; therefore, to avoid introducing further variability in the data, only BioVoice acoustic features were used at this study stage.

2.3. Machine learning experiments

Machine learning (ML) algorithms were implemented to conduct four separate classification experiments for female and male datasets. All signal processing steps were performed in MATLAB 2023a environment (The MathWorks, Inc., Natick, MS, USA).

All the involved subjects are adults (> 18 years old). However, the age varies widely (see Database subsection 2.1). Therefore, two experiments were conducted to test whether age could affect the results. The first concerns a classifier trained with the whole dataset (E1), while the second considers a subset of subjects aged 18–60 years only (E2). The outcome will define whether all successive ones will be carried out with the whole or the restricted database. E1 will also highlight if machine learning algorithms can distinguish between BLVF and UVFP patients with acoustic features only.

The third experiment (hereinafter E3) refines the previous ones, as its purpose is to further separate the BLVF class into three subclasses: nodules, polyps, and cysts [36]. This experiment was not performed for the M group as the sample size of the nodule patients is too small (2 subjects only, as shown in Table 1).

The fourth experiment (hereinafter E4) focuses on pre- and post-treatment observations to evaluate the efficiency of therapies. Acoustic recordings of the same patients before and after treatment were used. Before treatment, pathological patients are considered in one class, analogously to the post-treatment ones. Moreover, it has been evaluated if the models' errors in the post-treatment class belong to a specific pre-treatment pathology of the subjects (i.e. if the post-treatment error referred to a nodule, polyp, cyst, or vocal fold paralysis). This further analysis evaluated which pathology produced the highest error in the discrimination between pre and post-treatment. In E4, the identification tag of each misclassified recording was also stored to investigate a possible agreement between the AI model's outcome and the subjects' clinical voice assessment using the GRB scale [55,56]. A Mann-Whitney test with significance level $\alpha = 0.05$ was performed by considering G, R, and B indices between correctly classified and misclassified audio recordings.

The following supervised classifiers were considered and validated using a k-fold cross-validation framework ($k = 10$) [57]: SVM (with both linear and Gaussian kernel), Ensemble models (i.e., AdaBoost and RobustBoost using decision trees as weak learners, for more information readers are referred to [58,59]), and k- Nearest Neighbors algorithm (k-NN). Depending on the type of experiment (E1-E4) and the numerousness of each voice pathology condition across genders, Bayesian hyperparameters optimisation was used to maximise either the accuracy (ACC) or, in case of an imbalanced number of samples between groups, the Matthews Correlation Coefficient (MCC). Indeed, the literature demonstrated that the MCC represents a more reliable metric for micro-averaging correction in multi-class frameworks. [60]. Specifically, for E1 and E2, the ACC was used for hyperparameter tuning, whereas for E3 and E4, the MCC was implemented. During optimisation, the following hyperparameters were considered:

- For ensemble classifiers: number of learning cycles, minimum leaf size, the maximum number of splits, learn rate (only for AdaBoost), robust error goal (only for RobustBoost) and the split criterion (GDI or deviance). Specifically, the learning rate was evaluated in the range 0.001–1, the robust error goal between 0.001–0.49, the number of learning cycles between 10 and 300, the minimum leaf size between 1 and 50, the maximum number of splits between 1 and the size of the current training set for each experiment.
- For SVM: box constraint and kernel scale. The box constraint and kernel scale were evaluated in the 10–5 - 10^5 range. In the multiclass experiment (E3) the following coding designs were evaluated: one vs one, one vs all, binary complete (The MathWorks, Inc., Natick, MS, USA).
- For k-NN: number of nearest neighbours, distance metrics (Chebychev, Euclidean, Minkowski) and distance weights (equal, inverse, squared inverse). The number of nearest neighbours was evaluated in the 1–10 range.
- Moreover, for all models, the number of features, number of neighbours of ReliefF and misclassification costs were considered in the optimisation process. Specifically, the number of neighbours of ReliefF was evaluated in the range 1–15 and the misclassification costs in the range 1–10.

The number of iterations for Bayesian optimisation was set to 200. Each predictor was standardised with a z-score procedure during each cross-validation step for the training and validation sets, according to the current training statistics, avoiding possible data leakage.

Furthermore, to reduce the number of predictors, the following feature selection methods were applied: Pearson correlation, ReliefF [61], LASSO [62], and mrMR [63]. More precisely, before ReliefF, LASSO, and mrMR, the highly correlated predictors were removed from the training and validation set, retaining only those with an absolute Pearson correlation coefficient (PCC) < 0.8 in the training set [64]. When ranked by the considered feature selection methods, the number of features and the nearest neighbours for ReliefF were also used as

parameters during the Bayesian optimisation. The True Positive Rate (TPR) for each class and the area under the ROC curve (AUC) were also considered for binary classifiers. In contrast, for three-class problems, only the ACC, MCC, and AUC for each class were evaluated.

Models were trained using three different sets of features:

1. Features extracted from the sustained vowel /a/.
2. Features extracted from the Italian word /aiuole/
3. Merge of the two sets of features 1) and 2). In this case, subjects without one of the vocal tasks were removed.

This study addressed explainability by implementing transparent ML algorithms (k-NN) and predictor importance as a post-hoc technique for ensemble methods. The importance measure for each feature is computed with a two-step procedure [65]. First, changes in the node risk due to split are summed, where a change is defined as the difference between the risk for the parent node and the total risk for the two children. Then, the summation is divided by the total number of branch nodes. For example, if a tree splits a parent node (node 1) into two child nodes (nodes 2 and 3, respectively), then the algorithm increases the importance of the split predictor according to Equation (1):

$$imp = \frac{(R1 - R2 - R3)}{N_{branches}} \quad (1)$$

where imp is the feature importance, R_i is the node risk of node i, and $N_{branches}$ is the total number of branch nodes. A node risk represents a node error or node impurity weighted by the node probability. This is expressed by Eq. (2)

$$R_i = P_i \times E_i \quad (2)$$

where P_i is the node probability of node i, and E_i is either the node error (for a tree grown by minimising the twoing criterion) or node impurity (for a tree grown by minimising an impurity criterion, such as the Gini index or the deviance) of node i.

Finally, a statistical analysis was performed for the F and the M cohorts, considering the parameters the AI models retained after feature selection. A Mann-Whitney test was used, with a level of significance $\alpha = 0.05$. These tests were performed to evaluate if statistical differences exist between BLVF and UFVP. All the results related to the ML experiments and the statistical analysis are reported in Section 3.

3. Results

This section reports the results of the classification experiments described in Section 2.3. The acronym *fsm* denotes the feature selection methods by which these performances were obtained, while the number of features (nfeat) used to train the models are reported in brackets. In the tables, the mean and the standard deviation (μ and, respectively) obtained with the cross-validation of the methods are reported for each ML model. The results of the four experiments E1-E4 are presented in the subsections 3.1–3.4, respectively.

3.1. E1

In Table 4, the validation results obtained for the F and M datasets considering the binary classification between BLVF and UVFP concerning the E1 experiment are resumed in section 2.3). The performance considering the three feature sets (/a/, /aiuole/ and /a/+/aiuole/) is reported for both datasets. Only the models with the highest average accuracy in cross-validation are reported. For both datasets, an ensemble model obtained the highest performance. Fig. 1 reports the predictor importance for both. Table 5 displays which acoustic parameters present a significant difference between BLVF and UVFP in the age-unbalanced datasets for both female and male patients.

3.2. E2

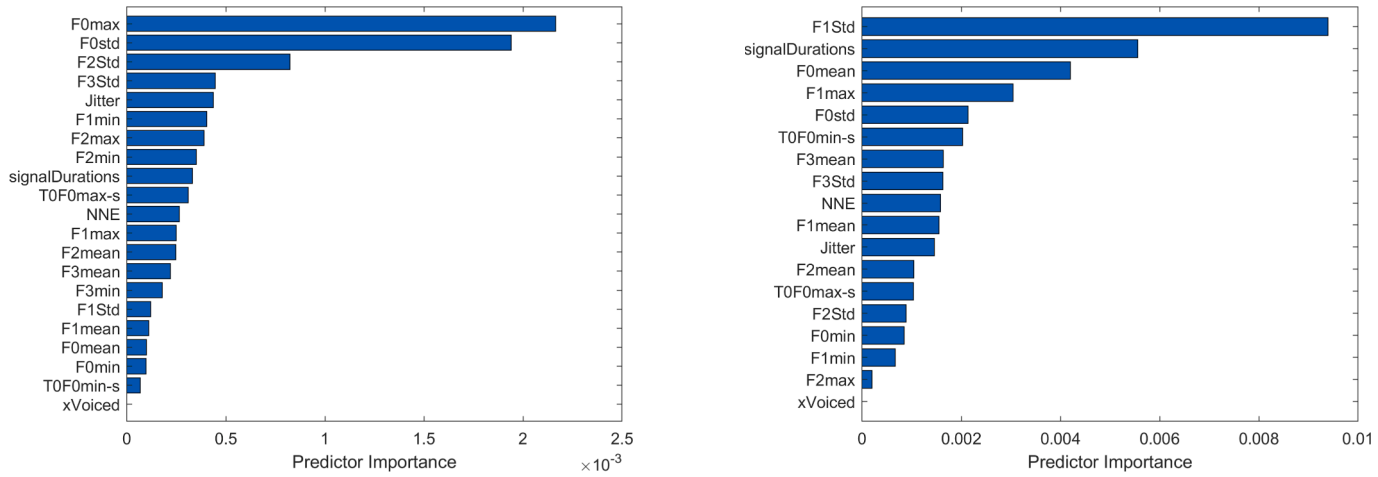
Table 6 reports the results related to the ML models considering an age-balanced dataset only for female patients diagnosed with BLVF and UVFP (E2). A Mann-Whitney test showed that, for adult females, age does not affect the difference between the two pathologies with $p - value = 0.1534$. The subset comprises 61 subjects with BLVF and 43 subjects with UVFP.

3.3. E3

Regarding the experiment on the classification of BLVF subclasses, i.e., nodules, polyps, and cysts, Table 7 reports the results obtained during the cross-validation (i.e., experiment E3, Section 2.3). Considering the unbalanced number of cases for each class, the MCC was used as a metric for the hyperparameter optimisation. Thus, the best model with the highest MCC (micro-averaging procedure for a multi-class problem) was

Table 4
Results of cross-validation (k-fold = 10) for the experiment E1 regarding the binary discrimination between BLVF and UVFP (indicating in the table their True Positive Rate with TPR^{BLVF} and TPR^{UVFP}, respectively). In this experiment, all the subjects were considered without pre-processing according to age. Mean and standard deviation ($\mu \pm \sigma$) statistics regarding the validation performance are reported. The metric used for hyperparameter tuning is highlighted in bold. The hyperparameters of the best models are also reported.

F dataset − E1							
Feature set	**ACC** (%)	TPR^{BLVF} (%)	TPR^{UVFP} (%)	MCC	AUC (%)	Model	fsm (numfeat)
/a/	76 ± 3	73 ± 4	77 ± 4	0.50 ± 0.06	79 ± 11	AdaBoost	Corr (21)
/aiuole/	74 ± 2	74 ± 4	72 ± 6	0.46 ± 0.06	75 ± 12	RobustBoost	Corr (19)
/a/ + /aiuole/	76 ± 3	76 ± 5	79 ± 4	0.55 ± 0.05	83 ± 9	AdaBoost	Corr(40)
Best F Model − E1							
Model Type	Learning rate		Learning cycles		Number of Leaves		Number of Splits
AdaBoost /a/	0.53		226		6		1
M dataset − E1							
Feature set	**ACC** (%)	TPR^{BLVF} (%)	TPR^{UVFP} (%)	MCC	AUC (%)	Model	fs (numfeat)
/a/	81 ± 2	84 ± 5	80 ± 5	0.64 ± 0.05	86 ± 14	RobustBoost	MRMR (18)
/aiuole/	73 ± 4	58 ± 10	80 ± 5	0.46 ± 0.07	67 ± 21	AdaBoost	Corr (20)
/a/ + /aiuole/	81 ± 3	82 ± 5	79 ± 5	0.62 ± 0.06	81 ± 14	RobustBoost	Corr (37)
Best M Model − E1							
Model Type	Robust error goal		Learning cycles		Number of Leaves		Number of Splits
RobustBoost /a/	0.44		152		2		80



(a) F dataset (b) M dataset

Fig. 1. Predictor Importance for the best ensemble models reported in Table 4 using Feature set /a/ both for F dataset (a) and M dataset (b).

Table 5
Statistical analysis results for acoustic parameters in E1.

F dataset – Significant features considered in the best model reported in Table 4			
Feature	p-value	BLVF – Median (IQR)	UVFP – Median (IQR)
F0 max /a/ [Hz]	e-06	249 (89)	302 (110)
F0 std /a/ [Hz]	e-04	8 (12)	14 (24)
F2 std /a/ [Hz]	0.001	59 (35)	75 (38)
Jitter /a/ [%]	0.007	0.98 (1.25)	1.42 (3.9)
F2 max /a/ [Hz]	0.01	1470 (267)	1522 (202)
Signal duration /a/ [s]	0.02	3.8 (2.1)	3.2 (1.9)
F1 max /a/ [Hz]	e-04	848 (177)	950 (252)
F1 std /a/ [Hz]	0.01	70 (36)	78 (52)
F0 mean /a/ [Hz]	e-04	189 (41)	209 (55)
M dataset – Significant features considered in the best model reported in Table 4			
Feature	p-value	BLVF – Median (IQR)	UVFP – Median (IQR)
F1 std /a/ [Hz]	e-07	38 (25)	66 (33)
Signal duration /a/ [s]	e-05	4.7 (2.3)	3.0 (2.1)
F1 max /a/ [Hz]	e-08	652 (137)	778 (105)
F0 std /a/ [Hz]	e-05	2.5 (5.2)	7.7 (27.9)
NNE /a/ [dB]	0.01	−21 (5)	−19 (13)
F1 mean /a/ [Hz]	0.01	547 (113)	598 (105)
Jitter /a/ [%]	e-04	0.84 (0.83)	3.02 (12.64)
F2 std /a/ [Hz]	e-05	40 (18)	57 (33)
F2 max /a/ [Hz]	0.001	1195 (115)	1271 (194)

selected. Only the details relative to the best models are reported in Table 7 for the F dataset and all of the feature sets considered. Moreover, considering the multi-class case, each subclass's AUC values are also reported. Finally, the M dataset was not considered for this analysis since the number of nodules recorded was too small (2 subjects).

3.4. E4

As reported in Section 2.3, the possibility of discriminating between pre- and post-treatment by ML models was evaluated. The results related to the former experiment (E4, Section 2.3) are reported in Table 8. In this case, in the "pre" class, both the BLVF and UVFP recordings were considered. As for the experiment reported in Table 7, the MCC metrics were used during the hyperparameters optimisation. Only the models with the highest average MCC obtained during cross-validation are reported. This analysis was performed for the F and M datasets, considering all the feature sets (/a/, /aiuole/, and /a/+/aiuole/). As for the results reported in Table 4, the predictor importance for both datasets is shown in Fig. 2.

To investigate possible relationships between AI model output and perceptual evaluation, only the models with the highest accuracy shown in Table 8 were considered for both datasets.

For the F group, a Mann-Whitney test highlighted that G, R, and B indices were significantly different between correctly classified and misclassified recordings of the /aiuole/ vocal task with $p - value = 0.02$, $p - value = 0.001$, and $p - value = 0.003$, respectively. Fig. 3 displays the distribution of the perceptual assessment indices.

Moreover, a significant difference was also found for the B index relative to the /a/ vocal task with $p–value = 0.01$. Fig. 4 shows the

Table 6
Results of cross-validation (k-fold = 10) for the experiment E2 regarding the binary discrimination between BLVF and UVFP on a balanced version of F dataset according to age. Mean (μ) and standard deviation (σ) statistics regarding the validation performance are reported. The metric used for hyperparameter tuning is highlighted in bold. The hyperparameters of the best model are also reported.

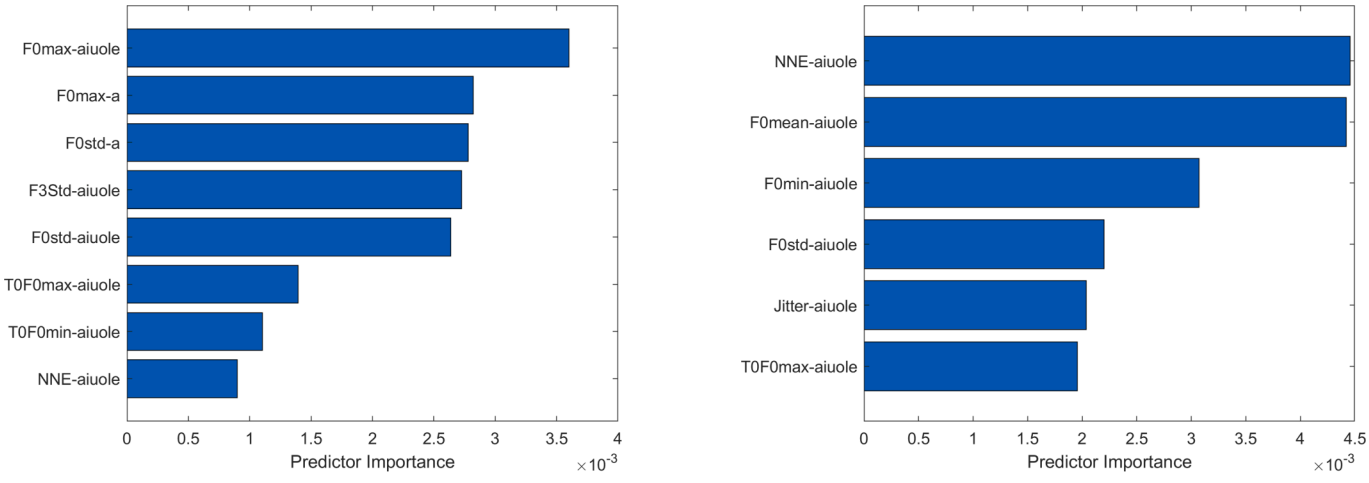
F dataset balanced according to the age – E2							
Feature set	**ACC** (%)	TPR^{BLVF} (%)	TPR^{UVFP} (%)	MCC	AUC (%)	Model	fsm (nfeat)
/a/	77 ± 4	80 ± 4	71 ± 8	0.52 ± 0.08	82 ± 11	AdaBoost	Corr (21)
/aiuole/	69 ± 2	78 ± 9	53 ± 11	0.41 ± 0.08	69 ± 19	RobustBoost	Corr (20)
/a/ + /aiuole/	72 ± 5	78 ± 6	68 ± 10	0.48 ± 0.09	77 ± 17	AdaBoost	Corr (41)
Best F Model – E2							
Model Type	Learning rate		Learning cycles		Number of Leaves		Number of Splits
AdaBoost /a/	0.65		294		2		2

Table 7
Results of cross-validation (k = 10) for the experiment E3, on the F dataset regarding the possibility to discriminate among the sub-classes representing BLVF: nodules (acronym = n), polyps (p) and cysts (c). For each class, the AUC values are reported. Mean (μ) and standard deviation (σ) statistics regarding the validation performance are reported. The metric used for hyperparameter tuning is highlighted in bold. The hyperparameters of the best model are also reported.

F dataset − E3							
Feature set	ACC (%)	MCC	AUC^n (%)	AUC^p (%)	AUC^c (%)	Model	fsm (nfeat)
/a/	51 ± 9	0.26 ± 0.07	57 ± 9	64 ± 5	41 ± 16	AdaBoost	Corr (22)
/aiuole/	60 ± 3	0.34 ± 0.06	70 ± 21	62 ± 5	64 ± 8	AdaBoost	Corr (19)
/a / + /aiuole/	54 ± 10	0.32 ± 0.10	73 ± 20	59 ± 17	64 ± 8	SVM	ReliefF(34)
Best F Model − E3							
Model Type	Learning rate		Learning cycles		Number of Leaves		Number of Splits
AdaBoost /aiuole/	0.51		264		9		26

Table 8
Cross-validation results related to binary pre vs. post evaluation (experiment E4). In this case, the pre-class merges the observations for the two groups considered (BLVF and UVFP), and the post-class represents the recording related to the post-treatment analysis for both groups. Mean (μ) and standard deviation (σ) statistics regarding the validation performance are reported for F and M datasets. The metric used for hyperparameter tuning is highlighted in bold. The hyperparameters of the best models are also reported.

F dataset − E4							
Feature set	ACC (%)	TPR^{pre} (%)	TPR^{post} (%)	**MCC**	AUC (%)	Model	fsm (nfeat)
/a/	76 ± 3	84 ± 4	57 ± 8	0.45 ± 0.05	78 ± 10	RobustBoost	ReliefF (9)
/aiuole/	75 ± 2	85 ± 4	53 ± 5	0.43 ± 0.05	75 ± 12	RobustBoost	Corr (20)
/a/ + /aiuole/	77 ± 4	83 ± 7	64 ± 6	0.52 ± 0.07	81 ± 11	RobustBoost	ReliefF(8)
Best F Model − E4							
Model Type		Learning rate	Learning cycles		Number of Leaves		Number of Splits
AdaBoost /a/+/aiuole/		0.42	125		24		1
M dataset − E4							
Feature set	ACC (%)	TPR^{pre} (%)	TPR^{post} (%)	**MCC**	AUC (%)	Model	fsm (nfeat)
/a/	71 ± 5	81 ± 3	55 ± 8	0.46 ± 0.05	72 ± 14	RobustBoost	MRMR (9)
/aiuole/	72 ± 3	76 ± 4	68 ± 7	0.48 ± 0.06	74 ± 13	RobustBoost	ReliefF (6)
/a/ + /aiuole/	70 ± 5	71 ± 4	65 ± 6	0.40 ± 0.09	66 ± 16	RobustBoost	MRMR (11)
Best M Model − E4							
Model type		Robust error goal	Learning cycles		Number of Leaves		Number of Splits
RobustBoost /aiuole/		0.30	218		4		58



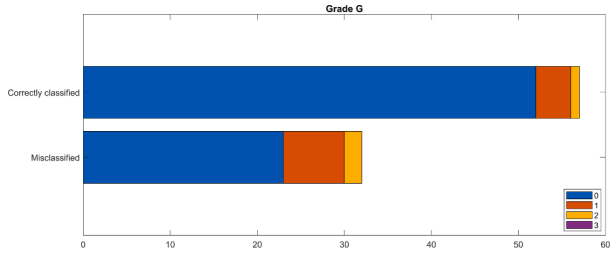
(a) F dataset (b) M dataset

Fig. 2. Predictor Importance for the best ensemble models reported in Table 8 regarding the pre- vs post-treatment analysis using feature set /a/+/aiuole/ for the F dataset (a), and feature set /aiuole/ for the M dataset (b).

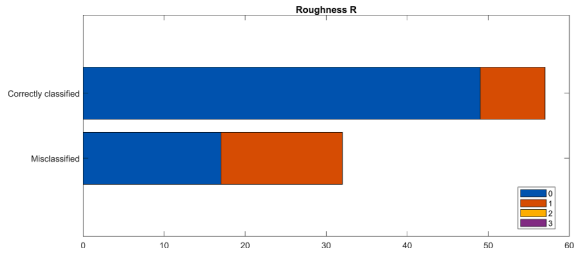
distribution of its values across correctly classified and misclassified post-treatment recordings.

For the M group, the misclassified recordings extracted from the /aiuole/ classifier did not present any significant difference in G, R, and B indices compared with correctly identified observations.

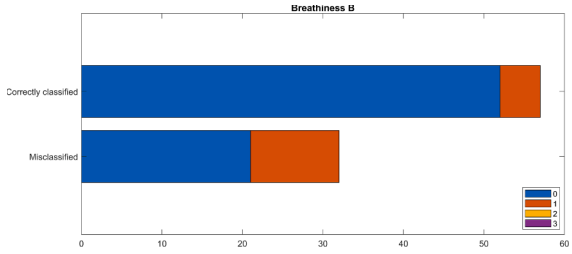
As reported in Table 8, the best models obtained the lowest TPR for



(a) G values distribution.



(b) R values distribution.



(c) B values distribution.

Fig. 3. Distribution of G, R, and B index values across correctly classified and misclassified observations from Table 8 best classifiers for the F dataset.

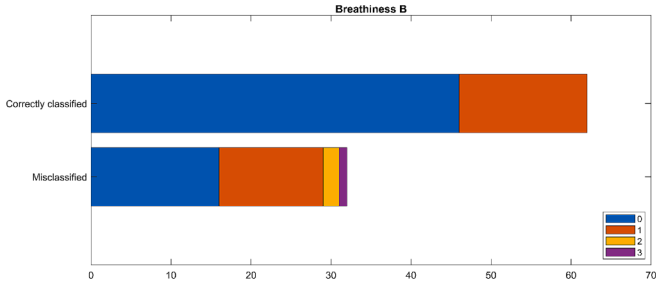


Fig. 4. Distribution of B index values across correctly classified and misclassified observations from Table 8 best classifiers for the M dataset.

the post-class for both the female and male datasets. The highest classification error was related to the UVFP in most cases. The best model of the Female dataset misclassified 27 observations. Their pre-treatment classes were UVFP (56 %, 18 cases), nodules (3 %, 1 case), polyps (25 %, 8 cases) and cysts (15 %, 5 cases). The best model of the Male dataset misclassified 19 cases. Their pre-treatment classes were UVFP (95 %, 18 cases) and polyps (5 %, 1 case).

Finally, Table 9 displays which acoustic parameters present a significant difference between pre- and post-treatment conditions for female and male patients.

Table 9
Statistical analysis results for acoustic parameters in E4.

F dataset – Significant features considered in the best model reported in Table 8			
Feature	p-value	PRE – Median (IQR)	POST – Median (IQR)
F0 max /aiuole/ [Hz]	e-09	286 (131)	223 (65)
F0 max /a/ [Hz]	e-10	270 (99)	212 (48)
F0 std /a/ [Hz]	e-10	11 (19)	4 (6)
F3 std /aiuole/ [Hz]	e-5	387 (137)	459 (161)
F0 std /aiuole/ [Hz]	e-10	28 (32)	17 (20)
M dataset – Significant features considered in the best model reported in Table 8			
Feature	p-value	PRE – Median (IQR)	POST – Median (IQR)
NNE /aiuole/ [dB]	0.03	−17 (7)	−18 (7)

4. Discussion

This paper proposes four machine learning experiments for voice pathology identification and characterisation based on acoustic features extracted from the sustained vowel /a/ and the Italian word /aiuole/. A focus on model interpretation is provided as well to:

- Evaluate the implicit role of age in the classification performance between E1 and E2.
- Highlight the predictors' importance and level of agreement with statistical analysis when comparing two (BLVF vs UVFP, E1) and three (nodules, polyps, cysts, E3) pathologies and the pre-post condition.
- Examine possible relationships between the output of the AI models and the GRB perceptual evaluation in the post-treatment condition (E4).
- Underline which *"pre"* classes were mostly misclassified in the pre-post experiment (E4).

E1 vs E2 classification comparison aims to highlight implicit bias introduced by age in the analysis, i.e., to understand how this factor may influence classification outcome. This study was performed on adult females only as some pathologies, especially nodules, present a strong gender-dependent prevalence [66]. Results in Tables 6 and 4 show that model accuracy does not improve when considering a subset where age does not significantly differ between BLVF and UVFP. Ageing influences vocal properties directly, supported by both statistical analysis [67–69] and AI-based methods [70], but the similar age distribution in our dataset and the chosen tasks might have reduced this effect. For /aiuole/, the accuracy changes from 69 % of balanced E2 to 74 % of unbalanced E1. Thus, an articulation task seems more sensitive to age than a sustained vowel. However, Spazzapan et al. [71] demonstrated that, for

female subjects, both sustained phonation and speech tasks highlighted statistical differences between age groups. Therefore, it seems implausible that one of the two tasks performs better in identifying UVFP than BLVF or vice versa due to a direct influence of age on vocal properties.

According to these results, we have performed further analyses for both genders on the unbalanced dataset only.

For the F dataset, the concurrent use of the acoustic features of /a/ and /aiuole/ did not bring relevant improvements. Indeed, as Table 4 illustrates, the accuracy is the same (76 %). The /a/+/aiuole/ classifier presents a higher AUC value than the /a/ one. However, such a result was obtained with a larger number of features (40 vs 21). Therefore, further examination and importance analysis of the predictor was performed on the /a/ classifier. The analysis of just the simple utterance /a/ , as well as reducing the number of parameters, could be helpful for clinicians. The acoustic features extracted from the sustained vowel gave a higher TPR for the UVFP class, confirming our previous outcome in [49]. This result can be of clinical relevance as a higher sensitivity in detecting vocal fold paralysis reduces the possibility that false-positive patients undergo long and expensive diagnostic procedures such as CT or MRI [72]. Unexpectedly, Fig. 1a shows that the most relevant parameters in distinguishing BLVF and UVFP are the fundamental frequency F0 and its instability F0std. Indeed, the vocal folds dynamic is quite different between these two pathologies. Moreover, the standard deviation of F2 was considered relevant by the AdaBoost model. Indeed, the second formant is related to the degree of constriction of the tongue movements [46]. This result may suggest otolaryngologists the use of imaging techniques such as ultrasonography or electromyography to track its activity to support and validate their diagnosis or treatment monitoring. The statistical analysis results in Table 5 show that the three most relevant parameters are also significant. UVFP presents higher values for the maximum F0, F0 std, and F2 std than BLVF, which may underline higher vocal fatigue and instability [73].

For the M dataset, E1 highlighted results similar to those of the F group. Considering the feature sets of both vocal tasks did not lead to performance improvements. The best classifier, in terms of accuracy, AUC, and number of parameters, was the RobustBoost using the feature set /a/. It is characterised by high TPR values for both BLVF and UVFP, 84 % and 80 %, respectively. Table 4 also shows a relevant result already found in [49]. For male adults, it seems that the sensitivity of voice pathology identification depends on the vocal task. Indeed, with the acoustic features of /a/, the TPR for benign lesions is higher than with the parameters of /aiuole/ (84 % vs. 80 %). In contrast, with the metrics of.

/aiuole/, the TPR for vocal fold paralysis is relevantly higher for the features of /a/ (80 % vs 58 %). A similar hypothesis was tested by Hosokawa et al. [74], showing that connected speech was better correlated with perceptual assessments than sustained phonation. However, this analysis was carried out between dysphonic and normophonic subjects rather than between different pathologies, as in the present work. Nevertheless, this result suggests that some vocal tasks might be more appropriate for diagnosing and detecting specific voice disorders. However, our classification outcome must be taken cautiously as the size of the M dataset is lower than the F one.

Interestingly, the most relevant acoustic feature for the E1 concerning the male group is the standard deviation of the first formant (Fig. 1b), which may be related to a difference in the instability of pharynx positioning and the degree of constriction [46]. Also, the overall signal duration, which can be considered an indicator of a patient's maximum phonation time, could represent a relevant clinical parameter to facilitate the diagnosis. Finally, the fundamental frequency appears to be the third most relevant parameter for distinguishing the pathologies considered in the M group compared to the F dataset. Overall, Fig. 1a and 1b show an agreement for both M and F datasets between acoustic parameters capable of separating BLVF and UVFP. However, the gender-specific relevance order in the predictor importance outcome suggests a distinct relationship between acoustic parameters and voice pathology that could be due to direct morphological differences in vocal fold and vocal tract structures. Finally, Table 5 shows that F1 std and signal duration significantly differ between BLVF and UVFP for the M dataset. The instability of the pharynx configuration is higher for the UVFP class. This result could still reflect greater vocal fatigue for patients diagnosed with such a pathology, whereas phonation capabilities are worse, likely determined by glottal insufficiency. Notably, the F0 variation between BLVF and UVFP was not significant. This could be due to the small size of the available male dataset. The E1 achieved the best performance with the acoustic features extracted from /a/.

To improve the generalisability of the results, more data should be acquired so that future research may also analyse if and how further acoustic parameters can improve this outcome and assess their relevance in the classification task. Moreover, additional observation should allow a correlation study with perceptual evaluation in the case of non-physiologically explainable features. Finally, the result of E1 opens the possibility of using such a model with other well-established voice pathology databases to test its robustness and to carry out cross-domain classification experiments. Indeed, sustained vowels are relatively independent of dialects and even language [1].

Due to the small M dataset sample size concerning nodules and cysts, E3 was performed for female patients only. Generally, Table 7 demonstrates a noticeably lower performance than the E1 and E2 classification outcomes. The highest accuracy was achieved for the AdaBoost /aiuole/ classifier with a value of 60 %, which aligns with the results of Marchese et al. [36]. The AUC for nodules (70 %) is higher than those for polyps and cysts (62 % and 64 %), which could indicate that the acoustic features of /aiuole/ may be sensitive to the dimensions of vocal fold masses. Indeed, Rostampourgonbki et al. [75] proved that fundamental frequency, jitter, and noise measures significantly differ between nodules, polyps, and cysts in female patients. However, this was not verified with predictor importance analysis due to the small sample size. Notably, even if the task, as well as the deployed data, is different, the best classifier in E3 presents a larger number of leaves and, mostly, a greater depth due to a higher number of splits than the best classifier in E1 (Table 7 vs Table 4, respectively). This result underlines that a more complex model was trained to perform this task, achieving moderate accuracy and MCC. Future research will address such problem.

In the overall pre-post comparison expressed in E4, the best model for the F dataset was an /a/+/aiuole/ RobustBoost classifier with a high accuracy of 77 %. Table 8 shows that the TPR for the pre-class was slightly worse and less stable with respect to the TPR of the /a/ classifier (83 $\pm$ 7 % vs 84 $\pm$ 4 %). However, the TPR for the post-class is higher (64 % vs 57 %). Therefore, the concurrent use of acoustic features extracted from different vocal tasks helped improve the sensitivity to detect post-treatment subjects. Even if TPR^{post} is relatively low, TPR^{pre} is promising. A high value means that, with the selected features, models tend to classify post-treatment recordings as *pre* treatment ones and not vice versa. Indeed, a patient diagnosed with a voice disorder is not considered a voice-recovered subject. Therefore, this model could reduce the possibility of under- or misdiagnoses. This could help clinicians monitor and evaluate the progress of therapies. Indeed, a post-treatment subject classified as a pre-treatment one could indicate a poor outcome of the intervention itself. Fig. 2a highlights that relevant acoustic features were extracted from the /aiuole/ vocal task, with the most important one being the maximum of the fundamental frequency. Hence, the subsequent articulation of four vocalic sounds could be a biomarker to detect a low voice quality after treatment. Moreover, vocal fold vibratory patterns expressed by F0-related metrics represent a meaningful characteristic that otolaryngologists should consider to assess the difference before and after therapy. Interestingly, all relevant metrics are related to the fundamental frequency except for the noise measure NNE extracted from /aiuole/. Therefore, E4 misclassified observations were further examined: firstly, it was investigated which class presented the highest error rate, then we performed a comparison

between classification errors and perceptual assessment of post-treatment voices to discover possible agreement between the results obtained with the proposed AI methods and the perceptual evaluation (Figs. 3 and 4). The first analysis highlighted that the most misclassified class was the UVFP. In our dataset, this pathology was surgically treated with a lipofilling procedure, which can reduce vocal fatigue with an easy and minimally invasive operation. Nonetheless, since it does not intervene on muscles or nerves, the motility of vocal folds is not entirely restored. Subsequently, voice quality might not be relevantly improved compared with the overall pre-treatment class. Fig. 3a seems to confirm this result, as a larger number of post-treatment recordings rated as more severe for the dysphonia grade G were misclassified as the pre-class. Similar considerations can be made for roughness and breathiness indexes (Fig. 3b and 3c). Moreover, the NNE extracted from /aiuole/ was relevant to distinguish pre- and post-treatment audio samples (Fig. 2a). Similarly, the R-value of misclassified observations was significantly higher with respect to correctly identified ones. It denotes a possible agreement between perceptual assessment and the classification outcome. It could also confirm that the NNE parameter is related to the R index of the GRB scale [76], similarly to the HNR [77]. Finally, Fig. 4 shows that breathiness B values for /a/ are higher for misclassified observations. Breathy voices typically characterise UVFP due to an incomplete glottic closure. Hence, the results of our model could underline a classification rule based on the UVFP pathophysiology.

In the future, it will be important to perform binary classification experiments to assess voice quality pre- and post- treatment for the BLVF and UVFP classes separately.

Table 8 shows that the best classifier for the M dataset was the RobustBoost model trained with the acoustic features of /aiuole/ only. Indeed, even if TPR^{pre} is lower with respect to the /a/ model (76 % vs. 81 %), the TPR^{post} is notably higher (68 % vs. 55 %). This means that fewer pre-treatment voice samples are misclassified as post-treatment ones. From Fig. 2b, the best acoustic metrics were NNE and fundamental frequency. Hence, noise level and vocal fold dynamics differ in the pre- and post-condition. No significant variation was identified in GRB values that might reveal an agreement between AI and perceptual assessment, possibly because the number of misclassified observations was too low to carry out a reliable statistical test. Indeed, when looking for the class that presented the highest error rate, it was found that the overall number of mistakes was equal to 19, out of which 95 % were UVFP. This confirms F dataset results: when considering BLVF and UVFP, models might face greater difficulty separating pre- and post-treatment recordings due to the different surgical operations adopted [78]. Such an outcome could be reflected and explained by the relevance of F0 as the surgical removal of vocal fold masses effectively restores their physiological motility. However, more data from a male cohort is needed to validate this outcome and hypotheses.

Differently from the predictor importance analysis displayed in Fig. 1, Fig. 2 shows that the pre/post-treatment models have considered alternative relevant features to separate the two classes, which, additionally, are also derived from the other vocal task /aiuole/, especially for the M dataset. This might indicate, again, time-intrinsic differences due to structural factors. Moreover, the difference between E1 and E4 predictor importance outcomes may suggest that the relevant features for voice pathology recognition are not fully identical to those for pre/post-treatment voice quality separation. Therefore, such acoustic parameters should not be used interchangeably for these classification experiments and, possibly, in clinical practice.

5. Conclusion

This paper proposes combining AI techniques and acoustic analysis to perform four experiments concerning the classification of BLVF and UVFP, the classification of nodules, polyps, and cysts, and the characterisation of pre- and post-treatment voice samples. The accuracy of the binary BLVF and UVFP classification experiment is high for both female and male datasets. Moreover, patients' age does not seem to bias the outcome, even if it is known that UVFP presents a higher disease onset than BLVF. Interestingly, pathology sensitivity is higher for UVFP than for BLVF female patients, which may be a clinically relevant result due to UVFP greater severity. On the other hand, sensitivity seems to be vocal task-dependent for M patients. Predictor importance highlighted a set of physiologically explainable parameters, with fundamental frequency and first formant variability being the most relevant and significant in the female and male databases. The three-class problem concerning the discrimination between nodules, polyps, and cysts obtained a rather low performance, achieved with /aiuole/ acoustic features, and it was performed for the female dataset only. In the future, more data is needed to understand whether this distinction between benign lesions is feasible and to reliably and robustly apply post-hoc explainability techniques. Finally, the pre/post-treatment experiment also showed a promising performance and a good agreement with the perceptual assessment, meaning that post-treatment samples misclassified as pre-treatment presented significantly worse GRB rates. However, the higher error rate for UVFP patients suggested that the pathologies considered in this study should be analysed separately to obtain a more comprehensive evaluation of surgical treatment efficiency.

In the future, this approach could be applied to a broader range of voice disorders to identify pathologies whose differential diagnosis still presents difficulties (e.g., spasmodic dysphonia and muscle tension dysphonia). Moreover, it could investigate the employment of additional acoustic parameters and the role of the time distance between surgery and voice recording for evaluating classification errors.

CRediT authorship contribution statement

Federico Calà: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lorenzo Frassineti:** Writing – review & editing, Visualization, Methodology, Formal analysis, Conceptualization. **Giovanna Cantarella:** Writing – review & editing, Supervision, Data curation, Conceptualization. **Giulia Buccichini:** Data curation. **Ludovica Battilocchi:** Data curation. **Claudia Manfredi:** Writing – review & editing, Supervision, Software, Conceptualization. **Antonio Lanatà:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] J.A. Gómez-García, L. Moro-Velázquez, J.I. Godino-Llorente, On the design of automatic voice condition analysis systems. Part I: review of concepts and an insight to the state of the art, Biomed. Signal Process. Control 51 (2019) 181–199, https://doi.org/10.1016/j.bspc.2018.12.024.
[2] L.O. Ramig, K. Verdolini, Treatment efficacy: voice disorders, J. Speech, Lang., Hearing Res. 41 (1) (1998) S101–S116, https://doi.org/10.1044/jslhr.4101.s101.
[3] R.H. Martins, H.A. do Amaral, E.L. Tavares, M.G. Martins, T.M. Gonçalves, N. H. Dias, Voice disorders: etiology and diagnosis, J. Voice 30 (6) (2016) 761–e1, https://doi.org/10.1016/j.jvoice.2015.09.017.
[4] J. Reid, P. Parmar, T. Lund, D.K. Aalto, C.C. Jeffery, Development of a machine-learning based voice disorder screening tool, Am. J. Otolaryngol. 43 (2) (2022) 103327, https://doi.org/10.1044/jslhr.4101.s101.
[5] M. González-Gamboa, H. Segura-Pujol, P.D. Oyarzún, S. Rojas, Are occupational voice disorders accurately measured? A systematic review of prevalence and

methodologies in schoolteachers to report voice disorders, J. Voice (2022), https://doi.org/10.1016/j.jvoice.2022.10.023.
[6] Z. Ali, G. Muhammad, M.F. Alhamid, An automatic health monitoring system for patients suffering from voice complications in smart cities, IEEE Access 5 (2017) 3900–3908, https://doi.org/10.1109/ACCESS.2017.2680467.
[7] J.D. Birkmeyer, A. Barnato, N. Birkmeyer, R. Bessler, J. Skinner, The impact of the COVID-19 pandemic on hospital admissions in the United States: study examines trends in US hospital admissions during the COVID-19 pandemic, Health Affairs 39 (11) (2020) 2010–2017, https://doi.org/10.1377/hlthaff.2020.00980.
[8] A.N. Omeroglu, H.MA. Mohammed, E.A. Oral, Multi-modal voice pathology detection architecture based on deep and handcrafted feature fusion, Eng. Sci. Tech., Int. J. 36 (2022) 101148, https://doi.org/10.1016/j.jestch.2022.101148.
[9] H.M. Mohammed, A.N. Omeroglu, E.A. Oral, MMHFNet: Multi-modal and multi-layer hybrid fusion network for voice pathology detection, Exp. Syst. Applic. 223 (2023), https://doi.org/10.1016/j.eswa.2023.119790.
[10] L. Verde, G. De Pietro, G. Sannino, Voice disorder identification by using machine learning techniques, IEEE Access 6 (2018) 16246–16255, https://doi.org/10.1109/ACCESS.2018.2816338.
[11] N.A.N. Za'im, F.T. Al-Dhief, M. Azman, M.R.M. Alsemawi, N.M.A. Abdul Latiff, M. Mat Baki, The accuracy of an online sequential extreme learning machine in detecting voice pathology using the malaysian voice pathology database, J. Otolaryngol.-Head Neck Surgery 52 (1) (2023) s40463–023, https://doi.org/10.1186/s40463-023-00661-6.
[12] A.-L. Hamdan, C. Jabbour, E. Khalifee, A. Ghanem, A. El Hage, Tolerance of patients using different approaches in laryngeal office-based procedures, J. Voice 37 (2) (2021) 263–267, https://doi.org/10.1016/j.jvoice.2020.12.009.
[13] H. Ghasemzadeh, M.K. Arjmandi, Toward optimum quantification of pathology-induced noises: an investigation of information missed by human auditory system, IEEE/ACM Trans. Audio Speech Lang. Process. 28 (2019) 519–528, https://doi.org/10.1109/TASLP.2019.2959222.
[14] I. Hidalgo-De la Guía, E. Garayzábal-Heinze, P. Gómez-Vilda, Voice characteristics in Smith-Magenis syndrome: an acoustic study of laryngeal biomechanics, Languages 5 (3) (2020) 31–43, https://doi.org/10.3390/languages5030031.
[15] J.I. Godino-Llorente, R. Fraile, N. Sáenz-Lechón, V. Osma-Ruiz, P. Gómez-Vilda, Automatic detection of voice impairments from text-dependent running speech, Biomed. Signal Process Control 4 (3) (2009) 176–182, https://doi.org/10.1016/j.bspc.2009.01.007.
[16] L. Frassineti, F. Calà, E. Sforza, R. Onesimo, C. Leoni, A. Lanatà, G. Zampino, C. Manfredi, Quantitative acoustical analysis of genetic syndromes in the number listing task, Biomed. Signal Process. Control 85 (2023) 104887, https://doi.org/10.1016/j.bspc.2023.104887.
[17] Z. Ali, M. Alsulaiman, I. Elamvazuthi, G. Muhammad, T.A. Mesallam, M. Farahat, K.H. Malki, Voice pathology detection based on the modified voice contour and SVM, Biol. Inspired Cognit. Archit. 15 (2016) 10–18, https://doi.org/10.1016/j.bica.2015.10.004.
[18] J.R. Duffy, et al., *Motor speech disorders: Substrates, differential diagnosis, and management*, Elsevier Health Sciences, 2012.
[19] F. Karlsson, L. Hartelius, How well does diadochokinetic task performance predict articulatory imprecision? differentiating individuals with parkinson's disease from control subjects, Folia Phoniatr. Logop. 71 (5–6) (2019) 251–260.
[20] A. Bayestehtashk, M. Asgari, I. Shafran, J. McNames, Fully automated assessment of the severity of parkinson's disease from speech, Comput. Speech Lang. 29 (1) (2015) 172–185.
[21] A.P. Vogel, M.L. Poole, H. Pemberton, M.W. Caverlé, F.M. Boonstra, E. Low, D. Darby, A. Brodtmann, Motor speech signature of behavioral variant frontotemporal dementia: refining the phenotype, Neurology 89 (8) (2017) 837–844.
[22] H. Kim, J. Jeon, Y.J. Han, Y. Joo, J. Lee, S. Lee, S. Im, Convolutional neural network classifies pathological voice change in laryngeal cancer with high accuracy, J. Clin. Med. 9 (11) (2020) 3415.
[23] Yu. Shih-Hau Fang, M.-J. Tsao, J.-Y. Chen, Y.-H. Lai, F.-C. Lin, C.-T. Wang, Detection of pathological voice using cepstrum vectors: a deep learning approach, J. Voice 33 (5) (2019) 634–641.
[24] I. Hammami, L. Salhi, S. Labidi, Voice pathologies classification and detection using EMD-DWT analysis based on higher order statistic features, Irbm 41 (3) (2020) 161–171, https://doi.org/10.1016/j.irbm.2019.11.004.
[25] R. Sharma, M. Kim, A. Gupta, Motor imagery classification in brain-machine interface with machine learning algorithms: classical approach to multi-layer perceptron model, Biomed. Signal Process. Control 71 (2022) 103101.
[26] A.M. Roy, An efficient multi-scale cnn model with intrinsic feature integration for motor imagery eeg subject classification in brain- machine interfaces, Biomed. Signal Process. Control 74 (2022) 103496.
[27] J. Sekar, P. Aruchamy, A novel approach for heart disease prediction using hybridized aith2o algorithm and sanfis classifier, Network Comput. Neural Syst. (2024) 1–39.
[28] S.A. Syed, M. Rashid, S. Hussain, H. Zahid, Comparative analysis of CNN and RNN for voice pathology detection, BioMed Res. Int. 2021 (1) (2021) 6635964.
[29] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, A. Othmani, Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech, Biomed. Signal Process. Control 71 (2022) 103107.
[30] M. Hireš, M. Gazda, P. Drotár, N.D. Pah, M.A. Motin, D.K. Kumar, Convolutional neural network ensemble for Parkinson's disease detection from voice recordings, Comput. Biol. Med. 141 (2022).
[31] N. Ganapathy, R. Swaminathan, T.M. Deserno, Deep learning on 1-d biosignals: a taxonomy-based survey, Yearbook of Medical Informatics 27 (01) (2018) 098–109.
[32] K. Degila, R. Errattahi, A. El Hannani, The UCD system for the 2018 FEMH voice data challenge, in: 2018 IEEE International Conference on Big Data (big Data), IEEE, 2018, pp. 5242–5246, https://doi.org/10.1109/BigData.2018.8622604.
[33] I. Miliaresi, A. Pikrakis, A modular deep learning architecture for voice pathology classification. IEEE Access (2023) 80465–80478, https://doi.org/10.1109/ACCESS.2023.3300795.
[34] P. Harar, Z. Galaz, J.B. Alonso-Hernandez, J. Mekyska, R. Burget, Z. Smekal, Towards robust voice pathology detection: investigation of supervised deep learning, gradient boosting, and anomaly detection approaches across four databases, Neural Comput. Applic. 32 (20) (2020) 15747–15757, https://doi.org/10.1007/s00521-018-3464-7.
[35] B. Woldert-Jokisz, S. Voice, Online, Database (2007).
[36] M.R. Marchese, F. Sensoli, S. Campagnini, M. Cianchetti, A. Nacci, F. Ursino, L. D'Alatri, J. Galli, M.C. Carrozza, G. Paludetti, A. Mannini, Artificial intelligence for the recognition of benign lesions of vocal folds from audio recordings, Acta Otorhinolaryngol. Ital. 43 (5) (2023) 317, https://doi.org/10.14639/0392-100X-N2309.
[37] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
[38] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery 9 (4) (2019) e1312.
[39] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (xai), IEEE Access 6 (2018) 52138–52160.
[40] N. Seedat, V. Aharonson, Y. Hamzany, Automated and interpretable m-health discrimination of vocal cord pathology enabled by machine learning, in: In *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, IEEE, 2020, pp. 1–6.
[41] W. Rahman, S. Lee, M.S. Islam, V.N. Antony, H. Ratnu, M.R. Ali, A.A. Mamun, E. Wagner, S. Jensen-Roberts, E. Waddell, T. Myers, Detecting parkinson disease using a web-based speech task: observational study, J. Med. Internet Res. 23 (10) (2021).
[42] F. Calà, L. Frassineti, C. Manfredi, P. Dejonckere, F. Messina, S. Barbieri, L. Pignataro, G. Cantarella, Machine learning assessment of spasmodic dysphonia based on acoustical and perceptual parameters, Bioengineering 10 (4) (2023) 426.
[43] B. Gutiérrez-Serafín, J. Andreu-Perez, H. Pérez-Espinosa, S. Paulmann, W. Ding, Toward assessment of human voice biomarkers of brain lesions through explainable deep learning, Biomed. Signal Process. Control 87 (2024) 105457.
[44] C.-H. Hung, S.-S. Wang, C.-T. Wang, S.-H. Fang, Using sincnet for learning pathological voice disorders, Sensors 22 (17) (2022) 6634.
[45] D. Zhao, Z. Qiu, Y. Jiang, X. Zhu, X. Zhang, Z. Tao, A depthwise separable cnn-based interpretable feature extraction network for automatic pathological voice detection, Biomed. Signal Process. Control 88 (2024) 105624.
[46] J.R. Deller Jr, J.H.S. Hansen, J.G. Proakis, *Discrete-time processing of speech signals*, Macmillan Publishing Co, New York, 1993.
[47] 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 Related Disorders 73 (2020) 23–30, https://doi.org/10.1016/j.parkreldis.2020.03.012.
[48] Y. Bensoussan, C. Park, M. Johns III, S. Brown, J. Pinto, J. Chang, M. Courey, A comparison of an artificial intelligence tool to fundamental frequency as an outcome measure in people seeking a more feminine voice, Laryngoscope 131 (11) (2021) 2567–2571, https://doi.org/10.1002/lary.29605.
[49] Federico Calà, Lorenzo Frassineti, Giovanna Cantarella, Ludovica Battilocchi, Giulia Buccichini, Antonio Lanatà, and Claudia Manfredi. AI techniques applied to acoustical features of paralytic dysphonia versus dysphonia due to benign vocal fold masses. In: Models and Analysis of Vocal Emissions for Biomedical Applications: 13th International Workshop, September, 12-13, 2023, pages 83–86. Firenze University Press, 2023.
[50] Lorenzo Frassineti, Alice Zucconi, Federico Calà, Elisabetta Sforza, Roberta Onesimo, Chiara Leoni, Mario Rigante, Claudia Manfredi, and Giuseppe Zampino. Analysis of vocal patterns as a diagnostic tool in patients with genetic syndromes. In: Models and Analysis of Vocal Emissions for Biomedical Applications: 12th International Workshop, December, 14-17, 2021, pages 83–86. Firenze University Press, 2021. doi: 10.36253/978-88-5518-449-6.
[51] B.L. Herrington-Hall, L. Lee, J.C. Stemple, K.R. Niemi, M.M. McHone, Description of laryngeal pathologies by age, sex, and occupation in a treatment-seeking sample, J. Speech Hear. Disord. 53 (1) (1988) 57–64, https://doi.org/10.1044/jshd.5301.57.
[52] Minoru Hirano. Clinical examination of voice. *Disorders of human communication*, 5: 1–99, 1981. ISSN 0173-170X.
[53] M.S. Morelli, S. Orlandi, C. Manfredi, Biovoice: A multipurpose tool for voice analysis, Biomedical Signal Processing and Control 64 (2021) 102302, https://doi.org/10.1016/j.bspc.2020.102302.
[54] Paul Boersma. Praat: doing phonetics by computer. *http://www. praat. org/*, 2007.
[55] Y. Maryn, N. Roy, M. De Bodt, P. Van Cauwenberge, P. Corthals, Acoustic measurement of overall voice quality: a meta-analysis, J. Acoust. Soc. Am. 126 (5) (2009) 2619–2634.
[56] P.H. Dejonckere, M. Remacle, E. Fresnel-Elbaz, V. Woisard, L. Crevier-Buchman, B. Millet, Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements, Revue De Laryngologie-Otologie-Rhinologie 117 (3) (1996) 219–224.

[57] T. Hastie, R. Tibshirani, J.H. Friedman, J.H. Friedman, The elements of statistical learning: data mining, inference, and prediction, Springer, New York, New York, 2009.
[58] S. González, S. García, J. Del Ser, L. Rokach, F. Herrera, A practical tutorial on bagging and boosting based ensembles for machine learning: algorithms, software tools, performance study, practical perspectives and opportunities, Information Fusion 64 (2020) 205–237.
[59] I.D. Mienye, Y. Sun, A survey of ensemble learning: concepts, algorithms, applications, and prospects, IEEE Access 10 (2022) 99129–99149.
[60] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics 21 (1) (2020) 1–13, https://doi.org/10.1186/s12864-019-6413-7.
[61] Kenji Kira and Larry A Rendell. A practical approach to feature selection. In *Machine learning proceedings 1992*, pages 249–256. Elsevier, 1992. doi: 10.1016/B978-1-55860-247-2.50037-1.
[62] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Royal Stat. Soc. Series b: Stat. Methodol. 58 (1) (1996) 267–288, https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
[63] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, J. Bioinformatics Computational Bio. 3 (02) (2005) 185–205, https://doi.org/10.1142/S0219720005001004.
[64] L. Frassineti, C. Manfredi, B. Olmi, A. Lanatà, A generalized linear model for an ECG-based neonatal seizure detector, in: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2021, pp. 471–474, https://doi.org/10.1109/EMBC46164.2021.9630841.
[65] Estimates of predictor importance for classification ensemble of decision trees - MATLAB. https://www.mathworks.com/help/stats/classreg.learning.classif.compactclassificationensemble.predictorimportance.html.
[66] A. Zhukhovitskaya, D. Battaglia, S.M. Khosla, T. Murry, L. Sulica, Gender and age in benign vocal fold lesions, Laryngoscope 125 (1) (2015) 191–196, https://doi.org/10.1002/lary.24911.
[67] A.O. Santos, J. Godoy, K. Silverio, A. Brasolotto, Vocal changes of men and women from different age decades: an analysis from 30 years of age, J. Voice 37 (6) (2021) 840–850, https://doi.org/10.1002/lary.24911.
[68] S. Taylor, C. Dromey, S.L. Nissen, K. Tanner, D. Eggett, K. Corbin-Lewis, Age-related changes in speech and voice: spectral and cepstral measures, J. Speech Lang. Hear. Res. 63 (3) (2020) 647–660, https://doi.org/10.1044/2019_JSLHR-19-00028.
[69] S. Gahl, R.H. Baayen, Twenty-eight years of vowels: Tracking phonetic variation through young to middle age adulthood, J. Phonetics 74 (2019) 42–54.
[70] L.A. Forero, E.C. Mendoza, M.M. Vellasco, M.A. Silva, J.A. Apolinário Jr, Classification of vocal aging using parameters extracted from the glottal signal, J. Voice 28 (5) (2014) 532–537.
[71] Evelyn Alves Spazzapan, Vanessa Moraes Cardoso, Eliana Maria Gradim Fabron, Larissa Cristina Berti, Alcione Ghedini Brasolotto, and Viviane Cristina de Castro Marino. Acoustic characteristics of healthy voices of adults: from young to middle age. In: *CoDAS*, volume 30. SciELO Brasil, 2018. doi: 10.1590/2317-1782/20182017225.
[72] S. Politano, F. Morell, K. Calamari, B. DeSilva, L. Matrka, Yield of imaging to evaluate unilateral vocal fold paralysis of unknown etiology, Laryngoscope 131 (8) (2021) 1840–1844, https://doi.org/10.1002/lary.29152.
[73] Y.A. Ras, M. Imam, M.M. El-Banna, N.H. Hamouda, Voice outcome following electrical stimulation-supported voice therapy in cases of unilateral vocal fold paralysis, Egyptian J. Otolaryngol. 32 (2016) 322–334, https://doi.org/10.4103/1012- 5574.192543.
[74] K. Hosokawa, T. Iwahashi, M. Iwahashi, S. Iwaki, C. Kato, M. Yoshida, D. Yoshida, I. Kitayama, M. Umatani, N. Matsushiro, et al., The significant influence of hoarseness levels in connected speech on the voice-related disability evaluated using voice handicap index-10, J. Voice 37 (2) (2020), https://doi.org/10.1016/j.jvoice.2020.11.024, 290.e7 290.e16.
[75] M. Rostampourgonbaki, N. Deghanpour, K. Kiakojouri, M. Dehghan, H. Gholinia, Acoustic voice measures in benign mass lesions, Iran. Rehabil. J. 20 (3) (2022) 363–368, https://doi.org/10.32598/irj.20.3.1596.1.
[76] Ben Barsties v. Latoszek, Youri Maryn, Ellen Gerrits, and Marc De Bodt. A meta-analysis: Acoustic measurement of roughness and breathiness. J. Speech, Lang., Hearing Res., 61(2):298–323, 2018. doi: 10.1044/2017_JSLHR-S-16-0188.
[77] T. Bhuta, L. Patrick, J.D. Garnett, Perceptual evaluation of voice quality and its correlation with acoustic measurements, J. Voice 18 (3) (2004) 299–304.
[78] G. Cantarella, R.F. Mazzola, M. Gaffuri, E. Iofrida, P. Biondetti, L.V. Forzenigo, L. Pignataro, S. Torretta, Structural fat grafting to improve outcomes of vocal folds' fat augmentation: long-term results, Otolaryngol.-Head Neck Surgery 158 (1) (2018) 135–143, https://doi.org/10.1177/0194599817739256.