applied sciences

MDPI

Article

Voice Disorder Multi-Class Classification for the Distinction of Parkinson's Disease and Adductor Spasmodic Dysphonia

Valerio Cesarini [1], Giovanni Saggio [1,*], Antonio Suppa [2,3], Francesco Asci [2], Antonio Pisani [4,5], Alessandra Calculli [4,5], Rayan Fayad [1], Mohamad Hajj-Hassan [6] and Giovanni Costantini [1]

1 Department of Electronic Engineering, University of Rome Tor Vergata, 00133 Roma, Italy
2 Department of Human Neurosciences, Sapienza University of Rome, 00185 Roma, Italy
3 IRCCS Neuromed Institute, Via Atinense, 18, 86077 Pozzilli, Italy
4 Department of Brain and Behavioral Sciences, University of Pavia, Via Agostino Bassi, 21, 27100 Pavia, Italy
5 IRCCS Mondino Foundation, Via Mondino, 2, 27100 Pavia, Italy
6 Department of Biomedical Engineering, Lebanese International University, Mazraa, Beirut 146404, Lebanon
* Correspondence: saggio@uniroma2.it

Abstract: Parkinson's Disease and Adductor-type Spasmodic Dysphonia are two neurological disorders that greatly decrease the quality of life of millions of patients worldwide. Despite this great diffusion, the related diagnoses are often performed empirically, while it could be relevant to count on objective measurable biomarkers, among which researchers have been considering features related to voice impairment that can be useful indicators but that can sometimes lead to confusion. Therefore, here, our purpose was aimed at developing a robust Machine Learning approach for multi-class classification based on 6373 voice features extracted from a convenient voice dataset made of the sustained vowel/e/ and an ad hoc selected Italian sentence, performed by 111 healthy subjects, 51 Parkinson's disease patients, and 60 dysphonic patients. Correlation, Information Gain, Gain Ratio, and Genetic Algorithm-based methodologies were compared for feature selection, to build subsets analyzed by means of Naïve Bayes, Random Forest, and Multi-Layer Perceptron classifiers, trained with a 10-fold cross-validation. As a result, spectral, cepstral, prosodic, and voicing-related features were assessed as the most relevant, the Genetic Algorithm performed as the most effective feature selector, while the adopted classifiers performed similarly. In particular, a Genetic Algorithm + Naïve Bayes approach brought one of the highest accuracies in multi-class voice analysis, being 95.70% for a sustained vowel and 99.46% for a sentence.

Keywords: Parkinson; dysphonia; voice; features; machine learning; classifier; AI

Citation: Cesarini, V.; Saggio, G.; Suppa, A.; Asci, F.; Pisani, A.; Calculli, A.; Fayad, R.; Hajj-Hassan, M.; Costantini, G. Voice Disorder Multi-Class Classification for the Distinction of Parkinson's Disease and Adductor Spasmodic Dysphonia. *Appl. Sci.* **2023**, *13*, 8562. https://doi.org/10.3390/app13158562

Academic Editor: Gino Iannace

Received: 23 June 2023
Revised: 22 July 2023
Accepted: 24 July 2023
Published: 25 July 2023

1. Introduction

Voice analysis has being evolving as a promising approach within the automatic assessment of a number of different pathologies due to non-invasiveness, ease of use, and the anytime anywhere accessibility of voice recording systems [1]. In particular, voice analysis can provide detection of some pathologies by measuring the deviation of selected acoustic parameters, becoming clinical biomarkers, with respect to healthy baselines [2], those pathologies including neurological diseases [3–5]. In particular, patients with Parkinson's Disease (PD) can manifest as a vocal tremor and speech difficulties in the early stage of disease [6], and subjects suffering from Adductor Spasmodic Dysphonia (ASD) can present with vocal cords spasming on the physiologic voice emission.

Objective diagnoses of these diseases are relevant to design appropriate treatment strategies, despite the currently adopted empirical-based methodologies that can suffer from subjectivity, being potentially biased by the knowhow of the examiner, and potentially suffering from inter- and intra-rater reliability issues. Within this frame, one of the most adopted rater scales for assessing the PD severity is the perceptual-based Unified Parkinson's Disease Rating Scale [7], whilst for ASD the diagnosis relies on perceptual evaluation

performed by ear by a voice analyst or phoniatrician. Even though the performances of specialist analysts enhance inter-analyst agreement, analysts often dispute each other's findings when rating symptom severity and importance [8].

PD and ASD can cause changes in voice quality, usually qualified as inducing hoarseness, strain, and reduced vocal control, as well as a rough and/or "wet" quality associated to the vocal emission. These shared symptoms, along with a subjective and unpredictable manifestation and severity of symptoms, make it difficult to differentiate between the two conditions based solely on vocal characteristics. Accurate PD and ASD diagnoses require expertise from healthcare professionals, such as neurologists and otolaryngologists. However, considering that currently there is a lack of widely accepted objective measures or biomarkers specific to one or the other disease, clinicians often act empirically, evaluating multiple factors that include, but are not limited to, a detailed medical history, a physical examination, and a perceptual analysis. With these challenges, a comprehensive evaluation by medical professionals experienced in both movement disorders and voice disorders is crucial for accurate diagnosis and monitoring of PD and ASD [9,10].

Conveniently, voice signals, which contain some of the most crucial information for detecting and/or isolating PD or ASD can be analyzed automatically and non-invasively by algorithmic means. Artificial Intelligence (AI) voice analysis relies on the extraction of specific features from the vocal signals, and on the classification performed by means of Machine Learning (ML) algorithms [5,11–13]. AI-enhanced voice analysis has been used with promising degrees of success for the identification or even staging of various voice-impairing pathological conditions, such as pulmonary diseases and COVID-19 [14,15], tremors and Parkinsonism [16], and even psychological assessments such as emotion recognition [17–19], and can be employed to perform a multi-class analysis for preliminarily identifying a patient's pathology.

This paper aims to explore AI methodologies for voice analysis based on acoustic features, as clinically relevant biomarkers, for the identification of PD vs. ASD vs. healthy conditions (gathered by control subjects). From validated PD and ASD vocal data we extracted a large number of acoustic features to avoid biasing from specific subsets. This is because, in general, voice analysis for PD detection often relies on some specific prosodic features like the fundamental frequency (F0), jitter, shimmer, Ssignal-to-noise ratio, or mel-frequency cepstral coefficients (MFCCs). However, such a subset of features alone may lead to underestimation and, moreover, there is no consistency or proven performance boost in using a certain subset with respect to another, as evidenced in a recent review [20]. As such, here, we extracted a large number of features employing a toolbox based on the INTERSPEECH 2016 feature set [21], which contains a vast amount of low-level descriptors (average, quartiles, delta coefficients, etc.) of the following domains: energy, spectral, Cepstrum [22], RASTA (RelAtive SpecTrAl) [23], voicing probability [24], F0, prosody (jitter, shimmer, Harmonic-to-Noise Ratio (HNR)), auditory loudness. This comprehensive feature set was algorithmically reduced, comparing several state-of-the-art methodologies of feature extraction. Then, three different classifiers were trained in order to look for the best combination.

2. Materials and Methods

2.1. Dataset

Voice recordings were performed in a double-walled, sound-attenuated room at the Policlinico Tor Vergata (PTV), an affiliate institution of the University of Tor Vergata (Rome-Italy). A digital audio recorder (ZOOM H5) was connected to a headset dynamic microphone (model WH20, by Shure Inc., Chicago, IL, USA) positioned 2–3 cm away from the mouth of the speaker, in a quiet environment with no significant room echo (measurable reverberation), no machinery noises, and/or any kind of static background, and there were no other speakers. Voice signals were sampled at 44.1 kHz, 16 bit resolution.

PD participants were a group of 51 PD patients (38 men and 13 women with an average age of 65 years), and a control group of 51 healthy subjects (HSs) (12 men and 39 women

with an average age of 63.7). ASD participants included 60 patients affected by ASD (9 men and 51 women with an average age 60.44 and 64.69 years, respectively), and a group of 60 age- and sex-matched HSs (15 men and 45 women with an average age of 60.73 and 57.76 years, respectively).

Each participant was asked to perform sustained emission of the vowel/e/ and to read loudly one selected Italian sentence ("Nella casa in riva al mare Maria vide tre cani bianchi e neri"). Vowel and sentence were repeated three times each, by applying a pitch considered as "normal" for each participant, i.e., without effort. The Italian sentence was chosen for its ability to activate the oral cavity while being uttered due to the presence and alternation of plosive and fricative consonant sounds.

2.2. Features

For each recording, 6373 features were extracted using the OpenSMILE tool (by Audeering®, Gilching, Germany, [25]) embedding the INTERSPEECH 2016 Computational Paralinguistics Challenge (ComParE 2016) [25] audio feature set, which contains the vast majority of relevant domains used in voice analysis, including time/energy, spectrum, cepstrum [22], RASTA (RelAtive SpecTrAl) [23], prosody, and perceptual features. First, a forward greedy step-wise filter with a correlation-based feature subset (CFS) evaluator was applied [26]. It is a supervised method based on the maximum-relevance, minimum-redundancy principle, computing a merit factor with correlation as a metric according to the following formula:

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

where k is the number of features in a subset S, $\overline{r_{fc}}$ is the average correlation between features and the class label, and $\overline{r_{ff}}$ is the average correlation between pairs of features in the subset.

After the CFS, a feature ranker was applied to further reduce the number of features, employing and comparing the following methods:

- Correlation: the features were ranked according to the value of their cross-correlation with the class.
- Information Gain (IG): measures the change in entropy, according to Shannon's definition, that a given dataset endures when it is the result of a split performed according to some criterion/threshold on each given feature. It is simply computed as a difference between post (conditioned) and prior entropy values [27,28].
- Gain Ratio (GR): normalized version of the IG, in order to tackle its potential sensitivity to the dataset cardinality, according to the following formula:

$$GainRatio(X) = \frac{IG(X)}{-\sum_{i=1}^{n} \frac{N(t_i)}{N_{tot}} * log_2 \frac{N(t_i)}{N_{tot}}} \quad (2)$$

  where IG is the IG for the feature X that presents n different values, and $N(t)$ is the number of occurrences of the value t. The denominator is defined as "split information". Note that, for continuous features, IG and GR are computed by sorting all the measured values and creating a corresponding number of splits, with each value acting as a below/above threshold [29].
- Genetic Algorithm (GA): a GA classifier was used as a wrapper, used on one feature at a time, to identify the best performing ones. A 10-fold cross-validation was employed to reduce data selection bias [30].

We started with a large-scale feature extraction (6373 features) rather than using preselected acoustic features to reduce selection bias and enable a thorough acoustical exploration of the voices by means of a statistically sound algorithmic reduction. This was in accordance with the principles behind the "Curse of Dimensionality", as defined by Taylor [31], which details how the number of features should ideally be comparable

to (if not less than) the cardinality of the dataset to avoid overfitting phenomena and loss of interpretability.

2.3. Classifiers

From the feature extraction and selection process, we obtained a total of four datasets (one for each selection method) per vocal task. With two vocal tasks (vowel/e/ and sentence), we have a grand total of eight datasets, used to train three classifiers:

- Naïve Bayes (NB).
- Random Forest (RF), with 100 iterations/bags created by sampling with repetition up to a dimension as big as the original set (100% bags).
- Multi-layer Perceptron (MLP), i.e., a fully connected artificial Neural Network with a number of hidden layers equal to the number of features + number of classes divided by two, trained with a learning rate of 0.3 and a momentum of 0.2.

With four selection methods (Correlation, IG, GR, and GA) and three classifiers (NB, RF, and MLP), there are twelve ML models for each vocal task, so 24 in total. We considered as the maximum test accuracy the "saturation value", that is, a value that does not meaningfully increase on increasing the number of features. For the wrapper method, the maximum accuracy result was guaranteed since it was the evaluator of the GA. Ten-fold cross-validation was used in all ML models: this leads to a data split of 90–10, chosen due to the dataset being limited in dimensions. Using 90% of the data allows us to retain a consistent (big enough) training set. On the other hand, although the test set consists of a small number of subjects, the ten cross-validation folds cover the whole dataset, thus training ten slightly different versions of each classifier, whose results on the test sets are averaged.

Receiving Operating Characteristic (ROC) curves are also produced, to assess the performance of the model over all its operating range for a given label or class. The Area Under ROC (AUROC) curve analysis of each class in every ML model developed is computed and used as a metric to assess the vocal test, classifiers, and feature selection methods.

2.4. Statistics

Statistical evaluation of experimental results is an essential part of the validation of new ML models. This is to infer if the new ML algorithm, feature selection, or vocal test, can provide a statistically significant improved performance with respect to others. Due to the nature of ML algorithms and datasets, we adopted non-parametric statistical methods [32,33].

In order to compare the ML algorithms, vocal tests, and feature selection methods against each other, we adopted the approaches of Iman and Davenport, Nemenyi, and Wilcoxon [32,34–36].

For comparing multiple classifiers, vocal tests, and feature selection methods, the Iman and Davenport test was used, based on Friedman's approach that is a non-parametric equivalent of ANOVA. It ranks by performance and tests the significance of multiple results. The hypothesis to test is that classifiers, vocal tests, or feature selection methods perform unequally. The null hypothesis that we are trying to reject is that these methods are equally performing. The Friedman's statistic is defined as follows:

$$\chi_F^2 = \frac{12 * N}{k * (k+1)} * \left[\sum_j R_j^2 - \frac{k * (k+1)^2}{4} \right] \tag{3}$$

distributed according to χ_F^2, with $k-1$ degrees of freedom, r_j^i being the rank of the j-th k algorithm on the i-th N dataset [32].

Iman and Davenport improved the conservatism of Friedman's statistic and developed a better one, defined as:

$$F_F = \frac{(N-1) * \chi_F^2}{N * (N-1) - \chi_F^2} \tag{4}$$

which is distributed according to the *F*-distribution with $k-1$ and $(k-1)(N-1)$ degrees of freedom [32,35].

For dual testing, the Wilcoxon Signed Rank Test, the Paired t-test, and Nemenyi's Test were used. Nemenyi's Test was the post hoc test following the Iman and Davenport test if the null hypothesis is rejected on comparing classifiers against each other. Nemenyi's test is defined as a critical difference as follows:

$$CD = q_{alpha} * \sqrt{\frac{k(k+1)}{6N}} \quad (5)$$

where alpha is the confidence level, k is the number of models, and N is the number of measurements. The obtained values of q (that change according to alpha) and k and are reported in Table 1. When the performances of the two classifiers, two vocal tests, or two feature selection methods under Friedman's statistic is larger than this CD, the performances are significantly different and unequal.

Table 1. Nemenyi's q-values with alpha = 0.05 and 0.10 according to the number of classifiers (k) considered in the statistical test.

Number of Classifiers	2	3	4	5	6	7	8	9	10
$q_{0.05}$	1.96	2.343	2.569	2.728	2.85	2.949	3.031	3.102	3.164
$q_{0.10}$	1.645	2.052	2.291	2.459	2.589	2.693	2.78	2.855	2.92

3. Results

The number of features needed to reach the maximum possible accuracies varied with respect to the vocal test, feature selection methods, and classifiers (Figure 1). The feature selection method that was a combination of the Correlation sentence MLP had the least number of features (52), whereas the Correlation Vowel/e/NB combination had the most (99).

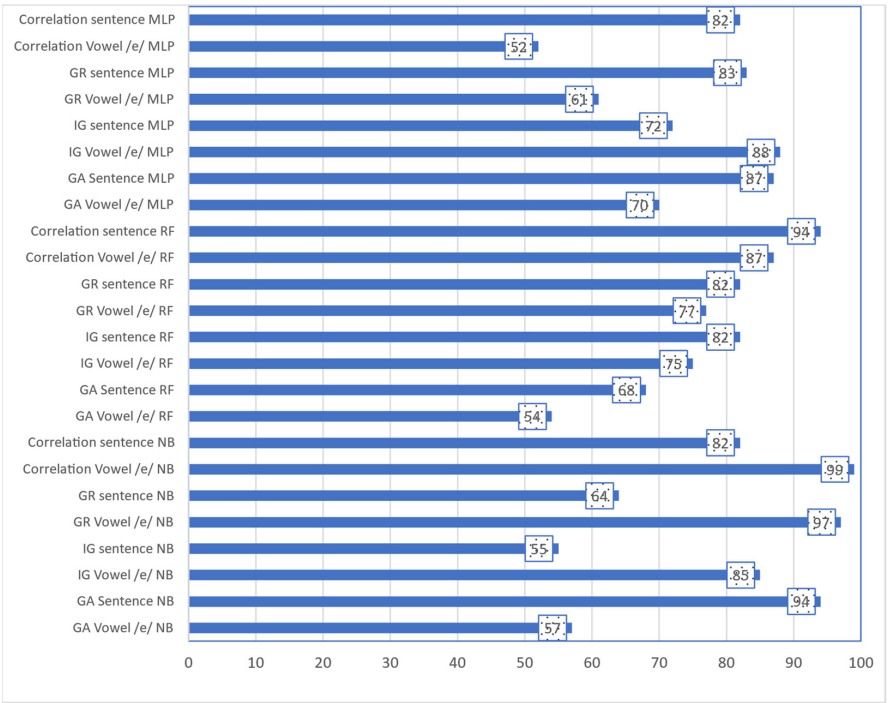


Figure 1. Optimal number of features to reach the best accuracy for each selection method and classifier. Abbreviations: MLP = Multi-layer Perceptron; GR = Gain Ratio; IG = Information Gain; GA = Genetic Algorithm; RF = Random Forest; NB = Naïve Bayes.

Figure 2 illustrates how the best accuracy was obtained using filter feature selection methods (Correlation, IG, GR). It demonstrates how the 10-fold cross-validation accuracy stopped increasing eventually while the number of selected features increased.

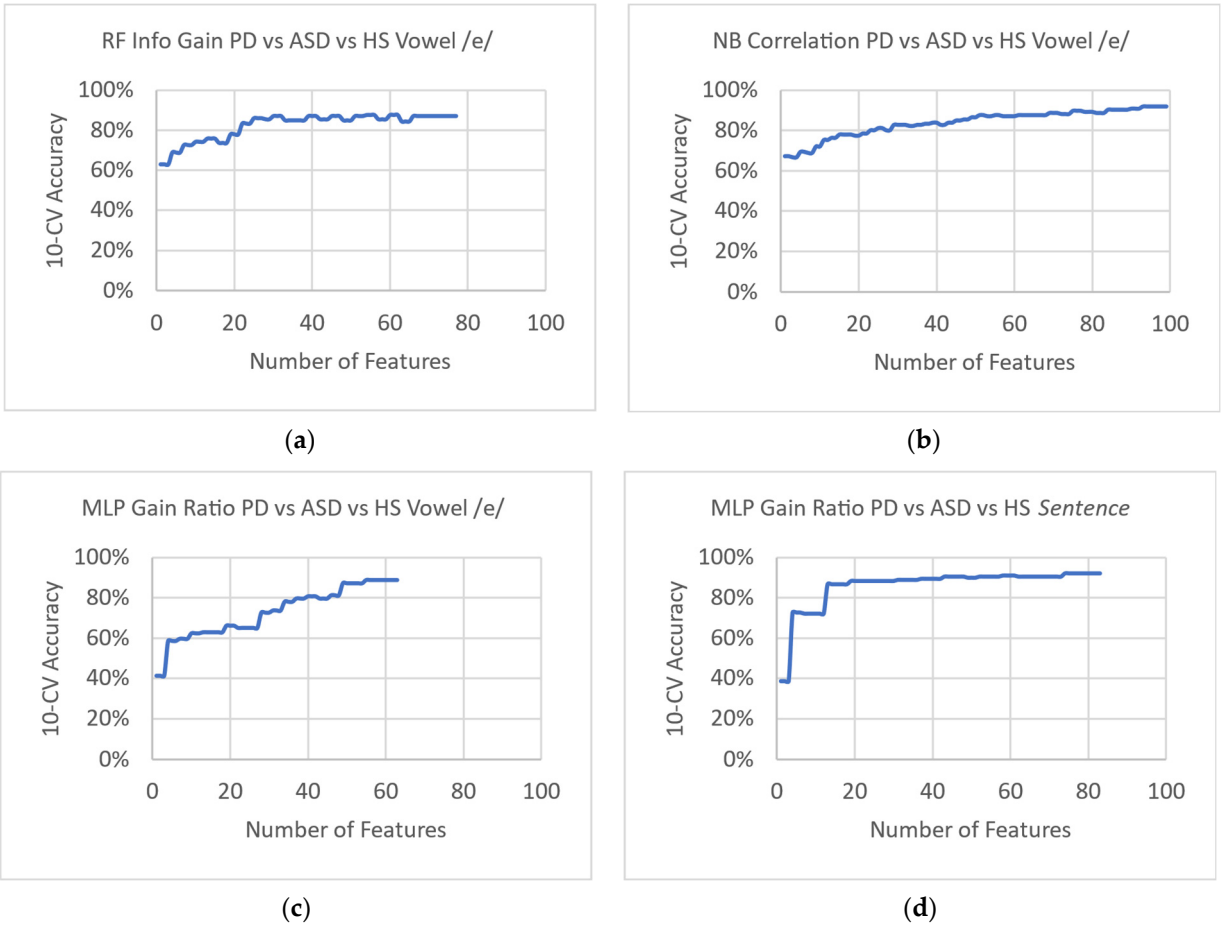


Figure 2. Example of accuracy saturation while the number of features for non-wrapper selectors increases. Graphs show the number of features versus classification accuracy for each multi-class (PD vs. ASD vs. HS) classification (**a**) RF classifier with IG selector, vowel/e/; (**b**) NB classifier with Correlation selector, vowel/e/; (**c**) MLP classifier with GR selector, vowel/e/; (**d**) MLP classifier with GR selector, sentence.

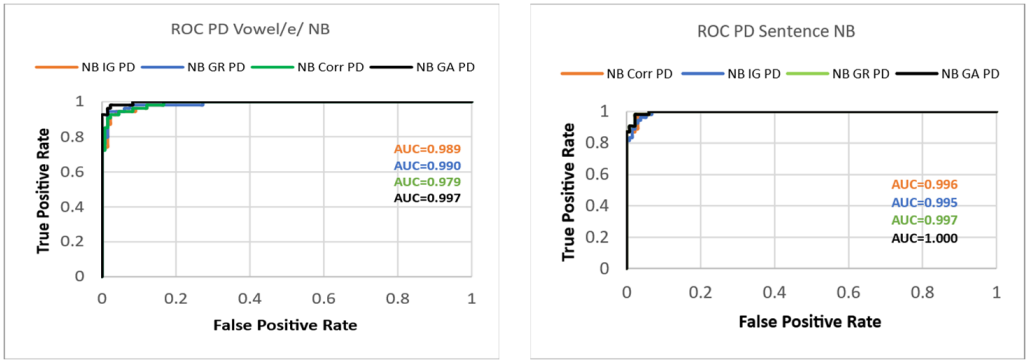
In the case of vocal test vowel/e/, the best results were found when classifying by NB and feature selecting by GA. The 10-fold cross-validation accuracy reached 95.70%. Furthermore, when doing AUROC analysis for this ML model, it is evident that the PD class had the highest AUROC (0.997), followed by HSs (0.994), and ASD (0.992). The lowest results were found when classifying by RF and feature selecting by GR. The 10-fold cross-validation accuracy reached was 87.10%. In addition, the HSs had the highest AUROC (0.975), followed by PD (0.974), and ASD (0.961).

Figure 3 shows an example of an obtained ROC curve.

The results from the MLP classifier were relatively lower than NB and better than RF. Selecting features with GA, the MLP 10-fold cross-validation accuracy reached 93.01%. In addition, the HSs had the highest AUROC (0.991), followed by PD (0.977), and ASD (0.937).

In the case of vocal test *sentence*, the best results were found when classifying by NB and feature selecting by GA. The ML model reached a 99.46% 10-fold cross-validation accuracy. In addition, by analyzing the ML model's ROC curves, it is evident that the PD class had the highest AUROC (1.00), followed by HSs (0.997), and ASD (0.997). The lowest results were found when classifying by RF and feature selecting by Correlation. The 10-fold

cross-validation accuracy reached was 87.63%. In addition, PD had the highest AUROC (0.991), followed by ASD (0.982), and HSs (0.973) (Figure 3). The results from the MLP classifier were relatively lower than NB and better than RF. Selecting features with GA, the MLP 10-fold cross-validation accuracy reached 98.39%. In addition, the PD had the highest AUROC (1.00), followed by HSs (0.981), and ASD (0.986). A further investigation into a selection of the top performing features, which allowed us to obtain the highest accuracy, obtained from the GA selection applied to the NB classifier on the sentence task, is presented in Table 2.



(**a**) (**b**)

Figure 3. Example of ROC curves for the NB classifier (all selectors), PD detection versus the other classes. (**a**) Vowel/e/ task; (**b**) sentence task.

Table 2. Top performing acoustic features, selected by the GA wrapper towards the NB classifier for the sentence vocal task. The names are according to the Compare 2016 nomenclature.

Features	Group of LLDS	LLD	Functionals	Group
mfcc_sma_de[3]_upleveltime25	Cepstral	MFCC	Up-Level Time 25%	Temporal
mfcc_sma_de[14]_meanFallingSlope	Cepstral	MFCC	Mean of Falling Slopes	Peaks
audSpec_Rfilt_sma_de[14]_quartile3	Prosodic	RASTA-style filtered auditory spectrum	Quartile 3	Percentiles
audSpec_Rfilt_sma_de[3]_quartile3	Prosodic	RASTA-style filtered auditory spectrum	Quartile 3	Percentiles
pcm_RMSenergy_sma_de_kurtosis	Prosodic	RMS Energy	Kurtosis	Moments
voicingFinalUnclipped_sma_de_centroid	Voice Quality	Voicing	Centroid	Temporal
pcm_fftMag_spectralSlope_sma_de_kurtosis	Spectral	Spectral Slope	Kurtosis	Moments
logHNR_sma_skewness	Voice Quality	HNR	Skewness	Moments
pcm_RMSenergy_sma_de_upleveltime25	Prosodic	RMS Energy	Up-Level Time 25%	Temporal
audSpec_Rfilt_sma[4]_lpc4	Prosodic	ZCR	Linear Prediction Coefficient 3	Modulation
pcm_zcr_sma_lpc3	Prosodic	ZCR	Linear Prediction Coefficient 4	Modulation
audSpec_Rfilt_sma[12]_segLenStddev	Prosodic	RASTA-style filtered auditory spectrum	Standard Deviation	Moments
pcm_fftMag_spectralEntropy_sma_peakRangeAbs	Spectral	Spectral Entropy	Amplitude Range of Peaks	Peaks

A summary of pathological speech (PD and ASD) detection results for all classifiers and selectors can be found in Table 3.

Table 3. Performance metrics for each classifier (NB, RF, MLP) and each feature selection method. The highest value (lowest for FP—False Positive) for each classifier, with respect to feature selectors, is stressed in bold. Abbreviations: NB = Naïve Bayes; RF = Random Forest; MLP = Multi-layer Perceptron; Corr. = Correlation feature selection; IG = Information Gain; GR = Gain Ration; GA = Genetic Algorithm; TP = True Positive; FP = False Positive; MCC = Matthew's Coefficient; ROC = Receiver-Operating Curve; PRC = Precision-Recall Curve.

Task	Classifier	Feature Selection	ACC	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Vowel/e/	NB	Corr.	91.94%	0.919	0.042	0.92	0.919	0.919	0.88	0.986	0.976
		IG	92.47%	0.925	0.038	0.925	0.925	0.924	0.887	0.984	0.974
		GR	93.55%	0.935	0.035	0.936	0.935	0.935	0.902	0.985	0.975
		GA	**95.70%**	**0.957**	**0.022**	**0.957**	**0.957**	**0.957**	**0.935**	**0.994**	**0.99**
	RF	Corr.	87.10%	0.871	0.067	0.872	0.871	0.87	0.806	0.97	0.953
		IG	88.17%	0.882	0.061	0.882	0.882	0.881	0.821	0.974	0.959
		GR	87.10%	0.871	0.068	0.871	0.871	0.87	0.806	0.97	0.95
		GA	**93.55%**	**0.935**	**0.033**	**0.936**	**0.935**	**0.935**	**0.903**	**0.983**	**0.972**
	MLP	Corr.	90.32%	0.903	0.05	0.903	0.903	0.902	0.856	0.965	0.944
		IG	89.25%	0.892	0.055	0.892	0.892	0.891	0.84	0.966	0.936
		GR	88.71%	0.887	0.059	0.887	0.887	0.886	0.831	0.961	0.935
		GA	**93.01%**	**0.93**	**0.036**	**0.931**	**0.93**	**0.93**	**0.896**	**0.969**	**0.949**
Sentence	NB	Corr.	93.01%	0.93	0.037	0.93	0.93	0.93	0.894	0.992	0.986
		IG	91.94%	0.919	0.042	0.919	0.919	0.919	0.877	0.985	0.966
		GR	93.01%	0.93	0.037	0.93	0.93	0.93	0.893	0.989	0.98
		GA	**99.46%**	**0.995**	**0.003**	**0.995**	**0.995**	**0.995**	**0.992**	**0.998**	**0.996**
	RF	Corr.	87.63%	0.876	0.069	0.88	0.876	0.876	0.813	0.981	0.966
		IG	89.78%	0.898	0.056	0.899	0.898	0.898	0.845	0.982	0.969
		GR	90.86%	0.909	0.048	0.909	0.909	0.909	0.861	0.982	0.968
		GA	**96.77%**	**0.968**	**0.018**	**0.969**	**0.968**	**0.968**	**0.951**	**0.991**	**0.984**
	MLP	Corr.	93.01%	0.93	0.037	0.932	0.93	0.93	0.895	0.982	0.969
		IG	91.40%	0.914	0.046	0.915	0.914	0.914	0.869	0.98	0.967
		GR	91.94%	0.919	0.042	0.92	0.919	0.919	0.878	0.982	0.97
		GA	**98.39%**	**0.984**	**0.009**	**0.984**	**0.984**	**0.984**	**0.975**	**0.988**	**0.978**

Figure 4 illustrates the respective acoustic groups of the selected features across all methods, vocal tests, and classifiers. The groups were spectral, cepstral, prosodic (energy related), and voice quality [37,38]. Spectral and prosodic groups were dominant across all feature selection methods whereas cepstral and voice quality groups varied in availability from one method to another. A further insight was made into the selection of top performing features by creating a radar chart, displayed in Figure 5. This chart was created using mean values of the PD and ASD features and normalized to its respective healthy ones. The chart shows an acoustic signature for both PD and ASD against the healthy voices.

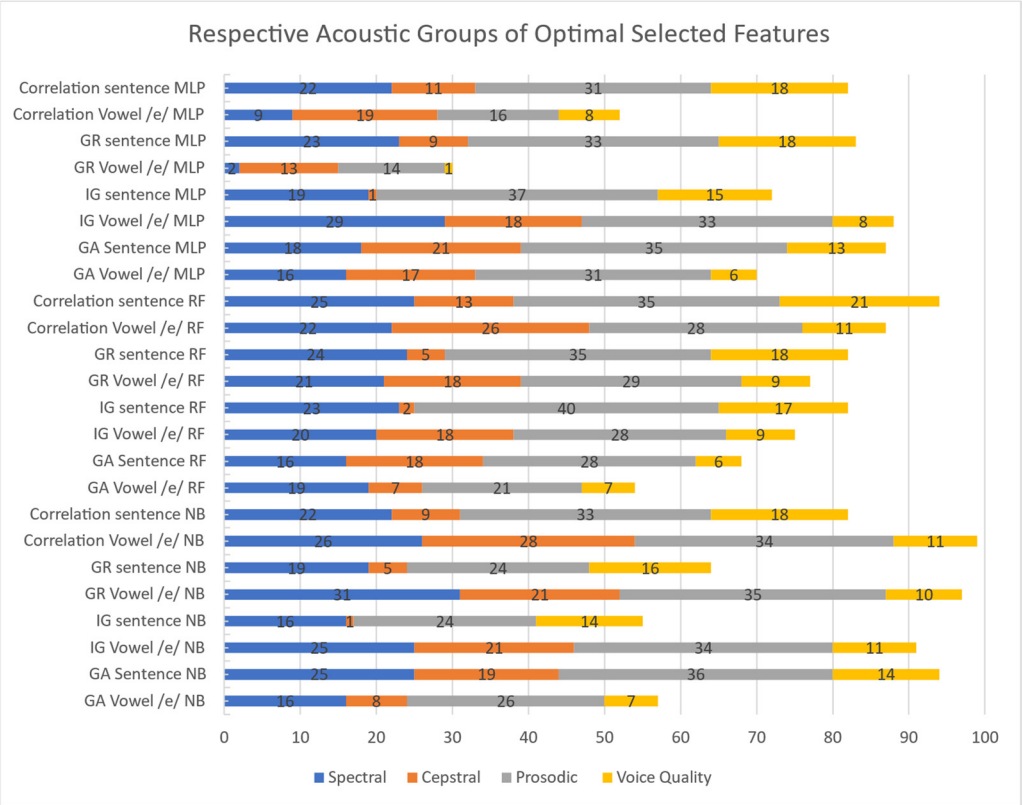


Figure 4. Respective acoustic group of optimal selected features: number of features retained after selection for each main feature domain.

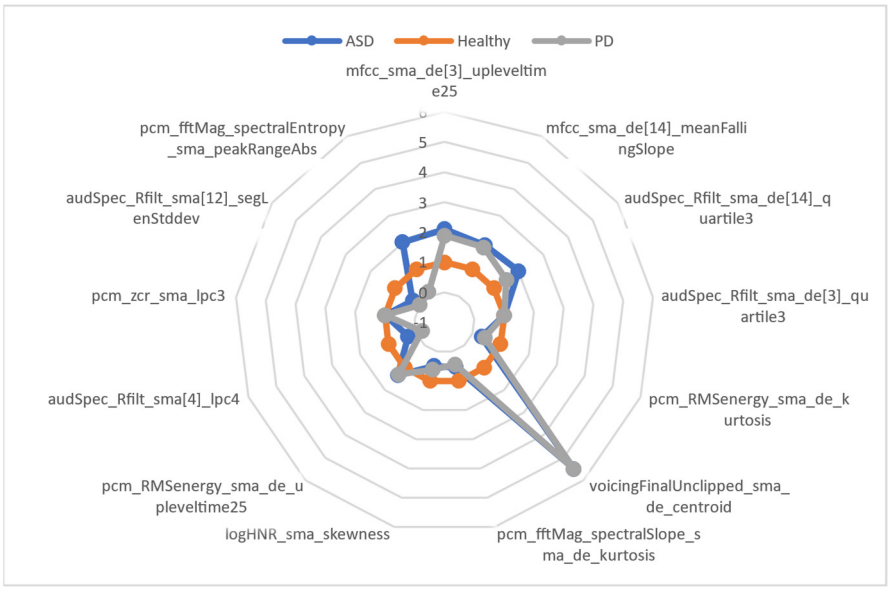


Figure 5. Radar chart displaying the different mean of the top performing features for each class (ASD, healthy, and PD), normalized by the healthy class (orange unit circle). Notice how ASD and PD behave similarly.

Table 4 reports the results of the statistical analyses performed on the obtained results which point out the statistical significance of the differences within classifiers and feature selectors. In the case of ML algorithms, the Iman and Davenport test was applied and the performance of classifiers was found to be statistically significant and not equivalent, with a p-value less than 0.0001. Comparing the classifiers against each other using Nemenyi's

Test, the NB's performance was found to be statistically significantly better than RF with a p-value less than 0.05. In addition, the NB's performance was tested using Wilcoxon's test against MLP and was found to be statistically significantly better, with a p-value of 0.01. Finally, MLP was tested against RF using Wilcoxon's and was found to be statistically significantly better with a p-value between 0.01 and 0.025. In the case of vocal tests, the Iman and Davenport test was used and the performance of both vocal tests was found to be statistically significant and not equivalent, with a p-value of 0.01. In addition, the performance of the sentence was better than vowel/e/. In the case of feature selection methods, the Iman and Davenport test was applied and the performance of these methods was found to be statistically significant and not equivalent, with a p-value of 0.04. Comparing the feature selection methods head to head, the GA's performance was found to be better than that of IG using Nemenyi's Test, with a p-value less than 0.05. Moreover, the performance of the GA was better than GR and Correlation using Wilcoxon's test, with a p-value of 0.05. However, when comparing, the performance of IG with GR, and Correlation with GR and IG, they were found to be statistically insignificant with p-values larger than 0.2. This suggests that their performance was similar.

Table 4. Results of all statistical inferences. Abbreviations: NB = Naïve Bayes, MLP = Multi-layer Perceptrion, RF = Random Forest.

Comparison	p-Value	Test	Null Hypothesis	Results
NB, MLP, RF	<0.0001	Iman and Davenport	The performances of all classifiers are equal	Performances are unequal
NB-RF	<0.05	Nemenyi	NB and RF are equal in performance	NB > RF
NB-MLP	0.01	Wilcoxon	NB and MLP are equal in performance	NB > MLP
RF-MLP	0.01–0.025	Wilcoxon	RF and MLP are equal in performance	MLP > RF
Sentence, vowel/e/	0.01	Iman and Davenport	Sentence and vowel/e/ have equal performances	Sentence performs better
Corr., IG, GR, GA	0.04	Iman and Davenport test	All feature selection methods are equal in performance	The performance of all feature selection methods is unequal
GA-IG	<0.05	Nemenyi test	GA and IG are equal in performance	GA > IG
GA-GR	0.05	Wilcoxon test	GA and GR are equal in performance	GA > GR
GA-Corr.	0.05	Wilcoxon test	GA and Correlation are equal in performance	GA > Correlation
IG-GR	>0.20	Wilcoxon test	IG and GR are equal in performance	IG and GR are similar in performance
Corr.-GR	>0.20	Wilcoxon test	Correlation and GR are equal in performance	Correlation and GR are similar in performance
Corr.-IG	>0.20	Wilcoxon test	Correlation and IG are equal in performance	Correlation and IG are similar in performance

4. Discussion

4.1. Literature Review

Voice features have been interesting subjects for researchers developing ML algorithms capable of classifying voice-related pathologies.

Within this frame, different databases were specifically adopted. As an example, Mykyska et al. [39] used The Massachusetts Eye and Ear Infirmary (MEEI) database that con-

sists of 53 healthy and 657 pathological speakers with different pathologies (e.g., ASD, conversion dysphonia, erythema, and hyperfunction), with the speaker uttering the vowel/a/. A second database was used, termed Príncipe de Asturias (PdA), which consists of 239 healthy and 200 pathological speakers with different pathologies (e.g., nodules, polyps, oedemas, and carcinomas), every speaker uttering the sustained vowel/a/. A third database was the so-called Czech Parkinsonian Speech Database (PARCZ), which consists of 52 healthy speakers and 57 PD patients who suffer from hypokinetic dysarthria, all speakers uttering the vowel/a/.

Feature extractions were done on assumed features for phonation, tongue movement, speech quality, segmental features, spectrum, wavelet decomposition, empirical mode decomposition, non-linear dynamics, and high level. A non-parametric Mann–Whitney U test was used for feature selection. Support Vector Machine (SVM) and RF classifiers were used for binary classification (pathological versus healthy). The accuracy of their results were 100% (MEEI), 82.1% $\pm$ 3.3% PdA, and 67.9% $\pm$ 6% (PARCZ), respectively.

Barche et al. [40] used the Saarbruecken Voice Database with 2000+ voice recordings sampled at 50 kHz, out of which 687 are collected from HSs (428 females and 259 males) and 1356 are collected from subjects (629 males and 727 females) with voice disorders. This database contains 71 different voice disorders. Each recording session consists of a German sentence and vowels of/a/,/i/, and/u. Only vowel vocal tests were used by the researchers who chose only six diseases from the dataset (spasmodic dysphonia, recurrent laryngeal nerve palsy, functional dysphonia, and psychogenic dysphonia, laryngitis, and leukoplakia). For feature extraction, they used the Interspeech Computational Paralinguistics Challenge features set (ComParE 2013), Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), MFCCs, Perceptual Linear Prediction (PLP), Glottal, Intonation, MFCC-Residual, and MFCC-ZFF (zero frequency filtering). A SVM classifier is used with five-fold cross-validation and binary classification. The best reported accuracy was 85.2%.

Verde et al. [41] used the Saarbruecken Voice Database. They chose all diseases and vowel/a/from the vocal tests. The features extracted were F0, jitter, shimmer, HNR, and MFCC. Feature selection was performed using correlation and IG. For classification they used SVM, Decision Tree (DT), Bayesian Classification (BC), Logistic Model Tree (LMT), and Instance-based Learning Algorithms (IBLA). The best reported result was an accuracy of 85.77% with the SVM binary (pathological versus healthy) classifier.

Alves et al. [42] used data from de novo patients recorded at the "Hospital das Clinicas" of the Medical School of Sao Paulo University. The database comprised 65 recordings for vowel/a/and have four groups: control (12 male and 8 female), Reinke's edema (2 males and 14 females), vocal nodules (2 males and 13 females), and neurologic diseases (7 males and 7 females). They used MFCC for feature extraction. For feature selection, they relied on statistics of the time series (voice segment) to search for a subset with greater discriminating power. SVM and K-Nearest Neighbors (KNN) were used for classification and cross-validation was applied. Their highest reported accuracy was 100% in all binary classifiers that paired the control group with one of the pathologies. In addition, classification accuracies of 99.08%, 98.86%, and 88.72% were obtained for the pairs Nodule/Neurological, Edema/Neurological, and Edema/Nodule, respectively. Al-Dhief et al. [43] used the Saarbruecken Voice Database and selected vowel/a/ recordings. For feature extraction the calculated MFCC. Classification was done using the Online Sequential Extreme Learning Machine (OSELM) algorithm. The best reported accuracy was 85%.

Gupta [44] used the Far Eastern Memorial Hospital (FEMH) voice disorder detection challenge database which includes 50 normal voice samples and 150 samples of common voice disorders, including vocal nodules, polyps, and cysts ("phonotrauma"), glottis neoplasm, and unilateral vocal paralysis. The voice samples are recordings for the speakers saying the vowel/a/. Feature extraction was done using MFCC. For classification, the author used the Long Short-Term Memory (LSTM) Recurrent Neural Network. The best reported result was 22% sensitivity, 97.1% specificity, and 56% unweighted average recall (UAR) [45]. Pham et al. used the FEMH database. Feature extraction was done using

MFCC. The ML classifiers used are SVM, RF, KNN, Gradient Boosting (GB), and Ensemble Learning (EL). The authors applied multi-class classification and the best reported accuracy was 68.48%.

Forero et al. [46] used a database from a speech therapist comprising two groups of patients, for a total of 12 speakers with vocal fold nodules, 8 speakers with vocal fold paralysis, and 11 speakers with normal voices. The vocal task was the sustained vowel/a/. They used inverse filtering to obtain the glottal signal and then extracted the following features: F0, jitter, shimmer, closing phase, opening phase, normalized amplitude quotient, amplitude quotient, closed quotient, open quotient, quasi-open quotient, speed quotient difference between harmonics H1 and H2, and harmonics richness factor. The classifiers used are artificial Neural Networks (ANN), SVM, and the Hidden Markov Model (HMM). The best reported result was 97.2% with multi-class classification [47]. Fang et al. used the FEMH database. MFCCs were the features extracted. Classifiers used were ANN, SVM, and GMM. The best reported accuracy was 99.32% in binary classification using ANN on the MEEI database. Table 5 reports a synoptic overview of the literature review of voice analysis for multi-pathology detection of dysphonic/neurodegenerative conditions.

A thorough review by Amato et al. [20] details more information regarding existing public datasets, feature toolboxes, trends in processing solutions, algorithms, and state of the art, concluding that voice analysis is still hindered by the problem of small and/or unclean datasets, and that acoustic features are often standardized into selective subsets that may contain only partial information, thus not allowing an extensive analysis.

Extensive research is done on the detection of PD [4,48–50] or ASD [5,13,51–53] from voices. Here, we expand our previous research developing a novel multi-class ML approach to distinguish de novo recordings of PD, ASD, and healthy control subjects using three different classifiers that inherently handle multi-class classification. This was obtained by using vigorous testing accuracy optimization, and cross- and statistical validations.

Table 5. Literature review. For more information about each study or dataset, please see the corresponding reference.

Study	Database	Pathology	Vocal Tasks	Feature Domains	Feature Selection	Classifier	Accuracy
Mekyska et al. [39]	MEEI, PdA, PARCZ	PD, ASD, conversion dysphonia, erythema, nodules, polyps, oedemas, carcinomas	/a/	Phonation, tongue movement, speech quality, spectrum, wavelet, EMD, non-linear dynamics	Mann–Whitney U test	SVM, RF	100%
Barche et al. [40]	SVD	Dysphonia (various), laryngeal nerve palsy, Laryngitis and Leukoplakia	/a/, /i/, /u/	eGeMAPS, MFCC, PLP, Glottal, Intonation, MFFC	N/A	SVM	85.20%
Verde et al. [41].	SVD	71 Pathologies	/a/	F0, Jitter, Shimmer, HNR, MFCC	Corr., IG	SVM, DT, BC, LMT, IBLA	85.77%
Alves et al. [42]	"Hospital das Clinicas"	Reinke's Edema, vocal nodules, neurologic diseases	/a/	MFCC	Statistics	SVM, KNN	100%
Al-Dhief et al. [43]	SVD	71 Pathologies	/a/	MFCC	N/A	OSELM	85%
Gupta [45]	FEMH	Vocal nodules, poylps, cysts, glottis neoplasm, unilateral vocal paralysis	/a/	MFCC	N/A	LSTM	56% UAR
Pham et al. [44]	FEMH	Vocal nodules, polyps, cysts, glottis neoplasm, unilateral vocal paralysis	/a/	MFCC	N/A	SVM, RF, KNN, GB, EL	68.48%
Forero et al. [47]	Speech therapist	Nodules and vocal paralysis	/a/	F0, jitter, shimmer	N/A	ANN, SVM, HMM	97.20%
Hemmerling et al. [49]	SVD	71 Pathologies	/a/, /i/, /u/	Various (28) + PCA	N/A	RF, Clustering	100%
Fang et al. [46]	FEMH	Vocal nodules, polyps, cysts, glottis neoplasm, unilateral vocal paralysis	/a/	MFCC	N/A	ANN, SVMM, GMM	99.32%
Ours	Custom (Tor Vergata)	PD, ASD	/e/+ sentence	Compare 2016 (6373)	Corr., IG, GR, GA	NB, RF, MLP	99.46%

4.2. Comments

In this study, we objectively detected PD, ASD, and HSs by means of advanced voice analysis based on ML, from our "de novo" dataset. Although we performed multi-class classification, we were able to achieve accuracies among the highest reported in the literature in the field of multiple pathologies using binary or multi-class classification, with NB providing the highest accuracy, and GA feature selection providing the best performing features. Feature selection played an instrumental role in achieving higher accuracies. It was also noteworthy to examine the acoustic groups that the selected features belong to, which points out how spectral, cepstral, prosodic, and voicing-related acoustic features are especially relevant for the detection and distinction of PD and ASD.

As evidenced by the aforementioned literature review and by Amato et al. [20], research works on voice classification differ in pathologies, languages, datasets, vocal tests, feature extraction and selection, accuracy, model validation, and statistical validation. Most importantly, they differ on whether the classification is binary or multi-class. Binary approaches will only classify whether the voice is pathological or healthy, making the ML model the least generalized and non-specific to any disease [39–43,45,46,49]. On the other hand, a multi-class approach is able to classify the nature and/or the severity of the pathology in the voice and, although requiring more data, it enables a more thorough analysis of voice impairment and its causes and evolutions.

Regarding the types of voice pathologies, some researchers used databases including a large number of pathologies [41,43,49]. This made the classification challenging, especially when the data are skewed towards certain pathologies, making generalization harder. We limited ourselves to two voice pathologies using a balance of data between ASD and PD to reduce bias and improve generalization. This is evident in the high true negative in our ML models. Furthermore, most researchers are re-using similar data, which limits the discovery of new patterns in the selected neurological diseases [39–41,43–47,49], whereas in our method we used de novo balanced data. Moreover, researchers used limited pre-selected features to extract, which narrows the investigation into the data and reduces data insights [41–47,49]. Whereas in our method we started from the complete 6373 feature set and then performed feature selection to find the most relevant and top performing features. In addition, some researchers did not validate their results [39,43–45], whereas we applied 10-fold cross validation to all of our models. Other researchers did not statistically validate the significance of their results [39–47,49] whereas we preferred adopting statistical tests to find significance across the vocal test, classifiers, and feature selection methods.

5. Conclusions

Several neurologic diseases are characterized by speech impairment that consistently worsens the quality of life of patients. The use of advanced voice analysis based on ML techniques in order to detect patients with PD and ASD would represent an advance in the field. In this study, we objectively detected and distinguished HSs from patients with PD and ASD by using multi-class classification based on a robust statistically validated approach. We focused on de novo balanced data, which were analyzed through feature extraction and selection processes. By applying this robust method, we reached a high ranking of 99.46% with a 10-fold cross-validation accuracy, using the NB classifier, GA feature selection, and sentence vocal test. Moreover, we explored the acoustic groups of the high performing data and observed the dominance of spectral and prosodic features. Exploring the top performing features further, we discovered an acoustic signature for PD and ASD compared to healthy voices. We consider our method to offer the advantage of ease access and thus it would help clinicians in improving clinical diagnosis also by expanding accessibility to a worldwide scale. Further research could be implemented by adding balanced data of additional voice disorders to further generalize the AI tool across multiple "dysphonia". Moreover, additional thorough statistical means exist for assessing the significance of the obtained results: the usage of the DeLong test for AUROC is especially relevant and will be considered for future works. Our research was performed

using a vocal task based on connected speech and it was limited to Italian-speaking people. However, the vowel/e/ vocal test can be collated in the future with other similar research. Correlation with GR and IG was found to be statistically insignificant with *p*-value larger than 0.2, meaning they had a similar performance.

Author Contributions: Conceptualization, G.C., A.S., G.S. and M.H.-H.; methodology, V.C. and G.C.; software, R.F.; validation, V.C.; formal analysis, V.C., G.C. and G.S.; investigation, V.C.; resources, A.S., F.A., A.P., A.C. and G.S.; data curation, A.S., F.A., A.P. and A.C.; writing—original draft preparation, R.F.; writing—review and editing, V.C. and G.S.; visualization, V.C. and R.F.; supervision, V.C., G.C. and G.S.; project administration, G.C., G.S., A.S. and A.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The dataset used in this study cannot be shared at the time being, due to privacy reasons.

Acknowledgments: The authors would like to thank all hospital staff involved in the study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Asci, F.; Costantini, G.; Di Leo, P.; Zampogna, A.; Ruoppolo, G.; Berardelli, A.; Saggio, G.; Suppa, A. Machine-Learning Analysis of Voice Samples Recorded through Smartphones: The Combined Effect of Ageing and Gender. *Sensors* **2020**, *20*, 5022. [CrossRef]
2. Saggio, G.; Costantini, G. Worldwide Healthy Adult Voice Baseline Parameters: A Comprehensive Review. *J. Voice* **2022**, *36*, 637–649. [CrossRef]
3. König, A.; Satt, A.; Sorin, A.; Hoory, R.; Toledo-Ronen, O.; Derreumaux, A.; Manera, V.; Verhey, F.; Aalten, P.; Robert, P.H.; et al. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimers Dement. Diagn. Assess. Dis. Monit.* **2015**, *1*, 112–124. [CrossRef] [PubMed]
4. Almeida, J.S.; Filho, P.P.R.; Carneiro, T.; Wei, W.; Damaševičius, R.; Maskeliūnas, R.; de Albuquerque, V.H.C. Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques. *Pattern Recognit. Lett.* **2019**, *125*, 55–62. [CrossRef]
5. Costantini, G.; Di Leo, P.; Asci, F.; Zarezadeh, Z.; Marsili, L.; Errico, V.; Suppa, A.; Saggio, G. Machine Learning based Voice Analysis in Spasmodic Dysphonia: An Investigation of Most Relevant Features from Specific Vocal Tasks. In Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021), Online, 11–13 February 2021; pp. 103–113. [CrossRef]
6. Parkinson's Foundation. 10 Early Signs of Parkinson's Disease', Parkinson's Foundation. Available online: https://www.parkinson.org/understanding-parkinsons/10-early-warning-signs (accessed on 18 July 2020).
7. Goetz, C.G.; Fahn, S.; Martinez-Martin, P.; Poewe, W.; Sampaio, C.; Stebbins, G.T.; Stern, M.B.; Tilley, B.C.; Dodel, R.; Dubois, B.; et al. The MDS-sponsored Revision of the Unified Parkinson's Disease Rating Scale. Milwaukee: International Parkinson and Movement Disorder Society. 2019. Available online: https://www.movementdisorders.org/MDS/MDS-Rating-Scales/MDS-Unified-Parkinsons-Disease-Rating-Scale-MDS-UPDRS.htm (accessed on 20 July 2020).
8. Hoffman, M.R.; Jiang, J.J.; Rieves, A.L.; McElveen, K.A.B.; Ford, C.N. Differentiating between adductor and abductor spasmodic dysphonia using airflow interruption: Differentiating Between SD Subtypes. *Laryngoscope* **2009**, *119*, 1851–1855. [CrossRef]
9. Merati, A.L.; Abaza, M.; Altman, K.W.; Sulica, L.; Belamowicz, S.; Heman-Ackah, Y.D. Common Movement Disorders Affecting the Larynx: A Report from the Neurolaryngology Committee of the AAO-HNS. *Otolaryngol. Neck Surg.* **2005**, *133*, 654–665. [CrossRef]
10. Lopes, B.P.; das Graças, R.R.; Bassi, I.B.; Neto, A.L.d.R.; de Oliveira, J.B.; Cardoso, F.E.C.; Gama, A.C.C. Quality of life in voice: A study in Parkinson's disease and in adductor spasmodic dysphonia. *Rev. CEFAC* **2012**, *15*, 427–435. [CrossRef]
11. Jiang, F.; Jiang, Y.; Zhi, H.; Dong, Y.; Li, H.; Ma, S.; Wang, Y.; Dong, Q.; Shen, H.; Wang, Y. Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc. Neurol.* **2017**, *2*, 230–243. [CrossRef] [PubMed]
12. Asci, F.; Costantini, G.; Saggio, G.; Suppa, A. Fostering Voice Objective Analysis in Patients with Movement Disorders. *Mov. Disord.* **2021**, *36*, 1041. [CrossRef] [PubMed]
13. Suppa, A.; Asci, F.; Saggio, G.; Marsili, L.; Casali, D.; Zarezadeh, Z.; Ruoppolo, G.; Berardelli, A.; Costantini, G. Voice analysis in adductor spasmodic dysphonia: Objective diagnosis and response to botulinum toxin. *Park. Relat. Disord.* **2020**, *73*, 23–30. [CrossRef]

14. Robotti, C.; Costantini, G.; Saggio, G.; Cesarini, V.; Calastri, A.; Maiorano, E.; Piloni, D.; Perrone, T.; Sabatini, U.; Ferretti, V.V.; et al. Machine Learning-based Voice Assessment for the Detection of Positive and Recovered COVID-19 Patients. *J. Voice* 2021, *in press*. [CrossRef]
15. Costantini, G.; Cesarini, V.; Robotti, C.; Benazzo, M.; Pietrantonio, F.; Di Girolamo, S.; Pisani, A.; Canzi, P.; Mauramati, S.; Bertino, G.; et al. Deep learning and machine learning-based voice analysis for the detection of COVID-19: A proposal and comparison of architectures. *Knowl.-Based Syst.* **2022**, *253*, 109539. [CrossRef]
16. Costantini, G.; Cesarini, V.; Di Leo, P.; Amato, F.; Suppa, A.; Asci, F.; Pisani, A.; Calculli, A.; Saggio, G. Artificial Intelligence-Based Voice Assessment of Patients with Parkinson's Disease Off and On Treatment: Machine vs. Deep-Learning Comparison. *Sensors* **2023**, *23*, 2293. [CrossRef]
17. Costantini, G.; Parada-Cabaleiro, E.; Casali, D.; Cesarini, V. The Emotion Probe: On the Universality of Cross-Linguistic and Cross-Gender Speech Emotion Recognition via Machine Learning. *Sensors* **2022**, *22*, 2461. [CrossRef] [PubMed]
18. Costantini, G.; Cesarini, V.; Brenna, E. High-Level CNN and Machine Learning Methods for Speaker Recognition. *Sensors* **2023**, *23*, 3461. [CrossRef] [PubMed]
19. Carrón, J.; Campos-Roca, Y.; Madruga, M.; Pérez, C.J. A mobile-assisted voice condition analysis system for Parkinson's disease: Assessment of usability conditions. *Biomed. Eng. Online* **2021**, *20*, 114. [CrossRef]
20. Amato, F.; Saggio, G.; Cesarini, V.; Olmo, G.; Costantini, G. Machine learning- and statistical-based voice analysis of Parkinson's disease patients: A survey. *Expert Syst. Appl.* **2023**, *219*, 119651. [CrossRef]
21. Schuller, B.; Steidl, S.; Batliner, A.; Hirschberg, J.; Burgoon, J.K.; Baird, A.; Elkins, A.; Zhang, Y.; Coutinho, E.; Evanini, K. The INTERSPEECH 2016 computational paralinguistics challenge: 17th Annual Conference of the International Speech Communication Association, INTERSPEECH 2016. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, San Francisco, CA, USA, 8–12 September 2016. [CrossRef]
22. Bogert, B.P. The quefrency alanysis of time series for echoes; Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. *Time Ser. Anal.* **1963**, *15*, 209–243.
23. Hermansky, H.; Morgan, N. RASTA processing of speech. *IEEE Trans. Speech Audio Process.* **1994**, *2*, 578–589. [CrossRef]
24. Yeldener, S. EP 1163662 A4 20040616—Method of Determining the Voicing Probability of Speech Signals. Available online: https://data.epo.org/gpi/EP1163662A4 (accessed on 24 May 2022).
25. Eyben, F.; Schuller, B. openSMILE:): The Munich open-source large-scale multimedia feature extractor. *SIGMultimedia Rec.* **2015**, *6*, 4–13. [CrossRef]
26. Hall, M.A. Correlation-based Feature Selection for Machine Learning. Ph.D. Dissertation, The University of Waikato, Hamilton, New Zealand, 1999.
27. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
28. Yücelbaş, C. A new approach: Information gain algorithm-based k-nearest neighbors hybrid diagnostic system for Parkinson's disease. *Phys. Eng. Sci. Med.* **2021**, *44*, 511–524. [CrossRef] [PubMed]
29. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
30. Sastry, K.; Goldberg, D.; Kendall, G. Genetic Algorithms. In *Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques*; Burke, E.K., Kendall, G., Eds.; Springer: Boston, MA, USA, 2006; pp. 97–125. [CrossRef]
31. Taylor, C.R. Dynamic Programming and the Curses of Dimensionality. In *Applications of Dynamic Programming to Agricultural Decision Problems*; CRC Press: Boca Raton, FL, USA, 2019. [CrossRef]
32. Demšar, J.; Demsar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
33. Razali, N.M.; Wah, Y.B. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J. Stat. Model. Anal.* **2011**, *2*, 21–33.
34. Benavoli, A.; Corani, G.; Mangili, F. Should we really use post-hoc tests based on mean-ranks? *J. Mach. Learn. Res.* **2016**, *17*, 152–161.
35. Iman, R.L.; Davenport, J.M. Approximations of the critical region of the fbietkan statistic. *Commun. Stat. Theory Methods* **1980**, *9*, 571–595. [CrossRef]
36. Ruxton, G.D.; Neuhäuser, M.; Ruxton, G.D.; Neuhäuser, M. When should we use one-tailed hypothesis testing?: One-tailed hypothesis testing. *Methods Ecol. Evol.* **2010**, *1*, 114–117. [CrossRef]
37. Weninger, F.; Eyben, F.; Schuller, B.W.; Mortillaro, M.; Scherer, K.R. On the Acoustics of Emotion in Audio: What Speech, Music, and Sound have in Common. *Front. Psychol.* **2013**, *4*, 292. [CrossRef]
38. Schuller, B.; Steidl, S.; Batliner, A.; Epps, J.; Eyben, F.; Ringeval, F.; Marchi, E.; Zhang, Y. The Interspeech 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load. In Proceedings of the INTERSPEECH 2014, 5th Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014; pp. 427–431.
39. Mekyska, J.; Janousova, E.; Gomez-Vilda, P.; Smekal, Z.; Rektorova, I.; Eliasova, I.; Kostalova, M.; Mrackova, M.; Alonso-Hernandez, J.B.; Faundez-Zanuy, M.; et al. Robust and complex approach of pathological speech signal analysis. *Neurocomputing* **2015**, *167*, 94–111. [CrossRef]
40. Barche, P.; Gurugubelli, K.; Vuppala, A.K. Towards Automatic Assessment of Voice Disorders: A Clinical Approach. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020. [CrossRef]
41. Verde, L.; De Pietro, G.; Sannino, G. Voice Disorder Identification by Using Machine Learning Techniques. *IEEE Access* **2018**, *6*, 16246–16255. [CrossRef]

42. Alves, M.; Silva, G.; Bispo, B.C.; Dajer, M.E.; Rodrigues, P.M. Voice Disorders Detection Through Multiband Cepstral Features of Sustained Vowel. *J. Voice* **2021**, *37*, 322–331. [CrossRef]
43. Al-Dhief, F.T.; Latiff, N.M.A.; Malik, N.N.N.A.; Sabri, N.; Baki, M.M.; Albadr, M.A.A.; Abbas, A.F.; Hussein, Y.M.; Mohammed, M.A. Voice Pathology Detection Using Machine Learning Technique. In Proceedings of the 2020 IEEE 5th International Symposium on Telecommunication Technologies (ISTT), Shah Alam, Malaysia, 9–11 November 2020; pp. 99–104. [CrossRef]
44. Pham, M.; Lin, J.; Zhang, Y. Diagnosing Voice Disorder with Machine Learning. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 5263–5266. [CrossRef]
45. Gupta, V. Voice Disorder Detection Using Long Short Term Memory (LSTM) Model. *arXiv* **2018**, arXiv:1812.01779.
46. Fang, S.-H.; Tsao, Y.; Hsiao, M.-J.; Chen, J.-Y.; Lai, Y.-H.; Lin, F.-C.; Wang, C.-T. Detection of Pathological Voice Using Cepstrum Vectors: A Deep Learning Approach. *J. Voice* **2018**, *33*, 634–641. [CrossRef] [PubMed]
47. Forero, M.L.A.; Kohler, M.; Vellasco, M.M.; Cataldo, E. Analysis and Classification of Voice Pathologies Using Glottal Signal Parameters. *J. Voice* **2015**, *30*, 549–556. [CrossRef]
48. Aich, S.; Kim, H.-C.; Younga, K.; Hui, K.L.; Al-Absi, A.A.; Sain, M. A Supervised Machine Learning Approach using Different Feature Selection Techniques on Voice Datasets for Prediction of Parkinson's Disease. In Proceedings of the 2019 21st International Conference on Advanced Communication Technology (ICACT), Pyeongchang, Republic of Korea, 17–20 February 2019. [CrossRef]
49. Hemmerling, D.; Orozco-Arroyave, J.R.; Skalski, A.; Gajda, J.; Nöth, E. Automatic Detection of Parkinson's Disease Based on Modulated Vowels. In Proceedings of the Interspeech 2016, San Francisco, CA, USA, 8–12 September 2016. [CrossRef]
50. Jeancolas, L.; Benali, H.; Benkelfat, B.-E.; Mangone, G.; Corvol, J.-C.; Vidailhet, M.; Lehericy, S.; Petrovska-Delacretaz, D. Automatic detection of early stages of Parkinson's disease through acoustic voice analysis with mel-frequency cepstral coefficients. In Proceedings of the 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Fez, Morocco, 22–24 May 2017. [CrossRef]
51. Fayad, R.; Hajj-Hassan, M.; Constantini, G.; Zarazadeh, Z.; Errico, V.; Saggio, G.; Suppa, A.; Asci, F. Vocal Test Analysis for the Assessment of Adductor-type Spasmodic Dysphonia. In Proceedings of the 2021 Sixth International Conference on Advances in Biomedical Engineering (ICABME), Werdanyeh, Lebanon, 7–9 October 2021; pp. 167–170. [CrossRef]
52. Schlotthauer, G.; Torres, M.E.; Jackson-Menaldi, M.C. A Pattern Recognition Approach to Spasmodic Dysphonia and Muscle Tension Dysphonia Automatic Classification. *J. Voice* **2010**, *24*, 346–353. [CrossRef]
53. Powell, M.E.; Cancio, M.R.; Young, D.; Nock, W.; Abdelmessih, B.; Zeller, A.; Morales, I.P.; Zhang, P.; Garrett, C.G.; Schmidt, D.; et al. Decoding phonation with artificial intelligence (D e P AI): Proof of concept. *Laryngoscope Investig. Otolaryngol.* **2019**, *4*, 328–334. [CrossRef]