Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 2 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i02.50503

PAPER

Exploring Essential Acoustic Features for Early Parkinson's Disease Classification: A Machine Learning Study

Daniel Hilário da Silva[1,2](✉), Caio Tonus Ribeiro[1], Leandro Rodrigues da Silva Souza[1,3], José Renato Munari Nardo[1], Adriano Alves Pereira[1,4]

[1]Programa de Pós-Graduação em Engenharia Biomédica, Uberlândia, Brasil

[2]Instituto Federal Goiano – Campus Cristalina, Cristalina, Brasil

[3]Instituto Federal Goiano – Campus Rio Verde, Rio Verde, Brasil

[4]Faculdade de Engenharia Elétrica, Uberlândia, Brasil

daniel.hilario@ifgoiano.edu.br

ABSTRACT

Parkinson's disease (PD) is a neurological condition that affects approximately 10 million individuals globally and is ranked as the second most prevalent neurodegenerative condition after Alzheimer's disease. Vocal disorders can be identified in approximately 90% of PD patients in the early stages of the disease. In this study, 19 machine learning (ML) algorithms were applied to a database of voice recordings of healthy individuals and individuals with PD obtained from a public repository. Different feature selection (FS) and hyperparameter optimization techniques were applied to all models for training, testing, and validation data. Among the ML algorithms, support vector machine with radial kernel, Naïve Bayes (NB), and Gaussian process classifier (GPC) yielded promising results when considering all features. Linear discriminant analysis, K neighbors classifier (KNN), extra trees classifier, GPC, and NB demonstrated excellent performance on the testing data after employing FS techniques. Decision tree classifier, KNN, and GPC emerged as the top performers when applied to the validation dataset. Our findings, derived from an extensive and chronological review of studies utilizing the same dataset, which surpass previous benchmarks, provide a comprehensive understanding of ML's application in voice analysis to support accurate clinical decision-making.

KEYWORDS

feature selection (FS), machine learning (ML), medical diagnosis, Parkinson's disease (PD), voice analysis

1 INTRODUCTION

Parkinson's disease (PD) is a neurological disorder characterized by the typical clinical signs and symptoms of bradykinesia, tremor, postural instability, and muscle rigidity [1]. It affects approximately 10 million people worldwide, with global

da Silva, D.H., Ribeiro, C.T., da Silva Souza, L.R., Munari Nardo, J.R., Pereira, A.A. (2025). Exploring Essential Acoustic Features for Early Parkinson's Disease Classification: A Machine Learning Study. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(2), pp. 98–120. https://doi.org/10.3991/ijoe.v21i02.50503

Article submitted 2024-06-09. Revision uploaded 2024-11-22. Final acceptance 2024-11-23.

occurrence doubling in the past 25 years due to increased longevity and a longer duration of illness [1]. It is the second most common neurodegenerative disease after Alzheimer's disease [2]. The clinical presentation includes motor and non-motor symptoms, and the diagnosis is based on the clinical features [3]. PD was first described by James Parkinson in 1817 [4], [5] under the name "shaking palsy." However, it was only in the 19th century that Charcot gave credit to Parkinson by referring to the disease as PD, which occurs due to the degeneration of cells located in a region of the brain called the substantia nigra [6]. Notably, the prediction of chronic diseases plays a pivotal role in healthcare informatics, and it is crucial to plan preventive actions and effective treatment for the early diagnosis of diseases [7]. Non-motor symptoms, including autonomic dysfunction, cognitive or neurobehavioral abnormalities, sleep disturbances, and sensory abnormalities, are common and are usually observed in patients with PD [6].

Decision-making on the status of a patient and the positive diagnosis of the disease still occur clinically, considering the medical history and physical examination, which makes the assessment and classification of the individual's status susceptible to errors. The main risk factor for PD is age; consequently, considerable growth is expected in the coming years owing to the aging population [8], [9]. Therefore, the identification of new markers for PD diagnosis or the improvement of the accuracy of the currently available tools is required, mainly in the initial stage of the disease [2], which is believed to occur before the manifestation of motor symptoms, with loss or decrease of sense of smell, sleep disturbances, tremor, and slowness of movement [10].

Noninvasive speech tests have been explored as a new marker for PD since speech deterioration, classified as one of the motor symptoms of the condition, is consistently observed in approximately 90% of patients with PD presenting with some vocal disorder in the initial stages [2], [10]. A previous study [11] showed that patients with PD were identified using acoustic features extracted from a speech test [2] and that voice is an early biomarker for PD detection [12]. Given the advancements in technology and computational power, machine learning (ML) has emerged as a valuable tool for the early prediction of diseases [12]. ML models are built from a considerable amount of data that are trained using mathematical and statistical approaches, allowing the classification, grouping, and prediction of new data.

In this study, PD was diagnosed using real data available in a public repository together with ML models, which were created using 19 different algorithms. Of these, six models were selected for a more detailed analysis in the case where all variables were initially considered and for the condition where some variables were disregarded using techniques widely used in the literature for feature selection (FS). This study focused on the application of ML models to classify patients with PD, considering data from voice records [10], [13], using a comprehensive approach to ML models. Our approach is distinguished by a comparative analysis with twelve other studies that have also utilized the same dataset [11]. This extensive review and comparison bring a new point of view to our study, not only benchmarking our models against existing research but also highlighting our novel contributions in terms of FS and algorithm optimization. This contextual backdrop underscores the significance of our work, illustrating a marked advancement over previous methodologies and setting the stage for our detailed investigation. In Section 2, we describe the public repository from which the data were selected, how the acoustic variables were obtained, and we describe in detail the methods for creating the ML models

applied to all features and a subset of features after FS techniques. Section 3 the results of the study. In Section 4, we discuss the findings, and in Section 5, present the conclusions of this study.

2 MATERIALS AND METHODS

In this section, we describe the dataset used, the preprocessing steps that we applied, and the ML techniques adopted in the study. The goal is to provide a clear understanding of the methodology followed to classify PD based on voice features, detailing each stage from data collection to model evaluation.

2.1 Dataset

The dataset used in the experiments consists of features obtained from the "Parkinson Dataset with Replicated Acoustic Features Data Set," which was donated to the UCI ML repository by Naranjo et al. [11] in 2019. The Bioethical Committee of the University of Extremadura approved the study protocol. The dataset presents the speech signals of 80 individuals, including 44 acoustic features extracted from the voice recordings of 40 patients with PD and 40 from the control group. Data were collected from 80 individuals over 50 years of age, as shown in Table 1.

Table 1. Distribution of the participants of the study

	Healthy Individuals	PD Individuals	Total
Recording number	120	120	240
Participants	40	40	80

The mean age (± standard deviation) was 66.38 ± 8.38 for the control group and 69.58 ± 7.82 for PD patients. According to a previous report, patients with PD present with at least two of the following symptoms: resting tremors, bradykinesia, and rigidity [11]. In the dataset, 240 records were observed, considering only 80 individuals with 46 variables in total, 44 variables with data over the voice, and two variables with data on the gender and status of the individuals [11], [14]. Of the 80 participants, 48 were male (22 from the healthy group and 26 from the PD group) and 32 were female (18 from the healthy group and 14 from the PD group). The recordings were obtained by asking the participants to produce a specific speech sustaining the intonation of the vowel /a/ for at least 5 seconds and repeating it three times. Voice samples were obtained through digital recording performed at a sampling rate of 44.1 kHz and 16-bit resolution using the Audacity software version 2.0.5 [11].

Table 2 displays all variables, including the voice measures used in the experiments, adapted from [10], [12], [14], and [15]. The 'Status' variable defines the class and has a value of 0 for healthy subjects and 1 for PD. The voice recordings yielded a set of 44 acoustic features, which could be categorized into five groups: pitch, local perturbation in amplitude, noise characteristics, special envelope features (including MFCCs and Delta coefficients), and nonlinear measures.

Table 2. Variables used in the experiments

Parameter	Abbreviation
Pitch	Jitter_rel (%), Jitter_abs, Jitter_RAP, Jitter_PPQ
Amplitude local perturbation	Shim_loc, Shim_dB, Shim_APQ3, Shim_APQ5, Shim_APQ11
Harmonic-to-noise ratio	HNR05, HNR15, HNR25, HNR35, HNR38
MFCCs and Delta Coefficients	MFCC0, MFCC1, ..., MFCC12
	Delta0, Delta1, ..., Delta12
Non-linear	RPDE, DFA, PPE, GNE
Others	Gender and Status

2.2 Preprocessing techniques

In the data preprocessing, standard libraries of the Python programming language were used with functions from the Python language to simplify the data processing procedure. The data were normalized using a z-score transform, leaving them with a zero mean and unitary variance, analogous to the concepts described by Rehman et al. [16]. The data standardization process was used in this study because the dataset contained numerous features with varied scales. Therefore, this step ensures that all features have the same weight during the model learning process [12], [16]. Initially, each participant contributed three replicated recordings, adhering to the methodology commonly adopted in related works [2], [15], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26]. These recordings served as independent records, with each containing distinct vocal features, a practice well-established in the field, as summarized in Table 9.

During the preprocessing phase, we imported the dataset in CSV format, as made available in the repository.[1] Initially, we omitted columns deemed irrelevant to our study, such as "ID" and "Recording." This step ensured the exclusion of nonessential metadata, streamlining the dataset for analysis. After preprocessing, the dataset retained 240 records, for a total of 46 variables. These included 44 acoustic variables, one variable representing gender (Gender), and one target variable (Status). We checked for any missing values in the dataset and found none, ensuring completeness.

2.3 Framework for classification modeling

For supervised ML modeling, 19 algorithms widely reported in the literature for PD classification and available in the PyCaret library were adopted for a comprehensive approach.

The proposed ML framework uses 46 features as predictor variables (standardized data: zero mean, unit variance) and disease status (PD or HC) as response variables. The initial number of acoustic features was 44, which was reduced after the FS process, as shown in Figure 1.

[1] The dataset used in the experiments consists of features obtained from the "Parkinson Dataset with Replicated Acoustic Features Data Set," which was donated to the UCI Machine Learning Repository.

The voice features and the optimal desired number were also selected based on their contributions to the ML models. Python language and some libraries, including Scikit-learn and PyCaret, were used [27].

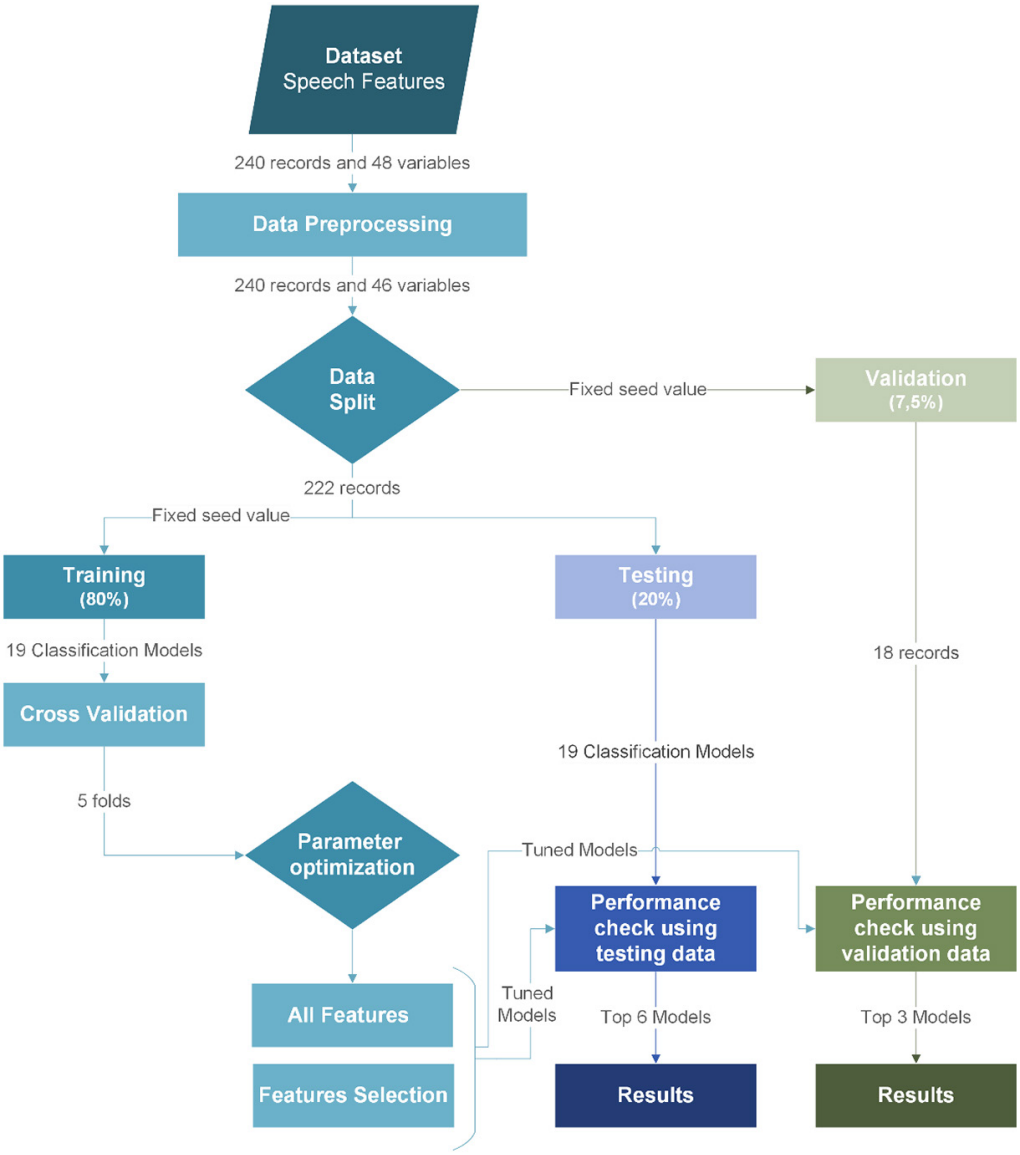

Fig. 1. Framework for ML modeling for PD classification with voice analysis

2.4 Data split and cross-validation

After the preprocessing stage, 7.5% of the dataset was separated into a validation subset (18 observations), while the remaining data (222 observations) was divided into two sets: training data and testing data, with proportions of 80% and 20%, respectively, as described in [16], similar to the division observed in [16], [25]. The training set was used to train the models and evaluate their performance using cross-validation (CV). The chosen strategy was stratified k-fold with a fold number set to 5, ensuring robustness in the evaluation process. Conversely, the test dataset was used to verify the model's generalization on unseen data. Table 3 presents the distributions of the three subsets.

Table 3. Distribution of the data in subsets for training, testing and validation

	Validation	Training (80%)	Testing (20%)
Number of records and variables	(r_{val}, v_{val})	(r_{tra}, v_{tra})	(r_{tes}, v_{tes})
	(18, 46)	(177, 46)	(45, 46)

The variables in Table 3 are defined as follows:

- r_{val} represents the number of records and v_{val} represents the number of variables used for the process of validation.
- r_{tra} represents the number of records and v_{tra} represents the number of variables used for training the models.
- r_{tes} represents the number of records and v_{tes} represents the number of variables used for testing the trained models.

The division of the dataset, as presented in Table 3 and visually represented in Figure 1, was pivotal in ensuring the robustness of our model evaluations. Initially, we segmented the dataset into three subsets: training, validation, and testing.

2.5 Voice feature selection techniques

For the first approach, the proposed structure for the ML models examined 46 features, 44 acoustic features, and one feature containing the individual's gender as a predictive variable since the disease state (PD or HC) was considered as the response variable. In this study, we used FS, a preprocessing phase in data science, to identify the key features of the problem. FS has several advantages, such as saving time for future data collection, understanding the causes of diseases, lowering computational costs, and having no degradation in performance [10]. A satisfactory determination of the variables during the FS phase can improve the accuracy of the classification algorithm [7], [10].

The FS process was initially performed on the 46 features of the problem to obtain a compact and effective model. Two FS algorithms, principal component analysis (PCA) [15], [20] and correlation feature selection (CFS) [28], were applied to the different classification methods. The algorithms then generated new feature subsets and classifications. The application of these techniques aims to determine the optimal number of features by considering their contribution to improving the accuracy of the classification algorithms. Figure 2 illustrates the FS process.

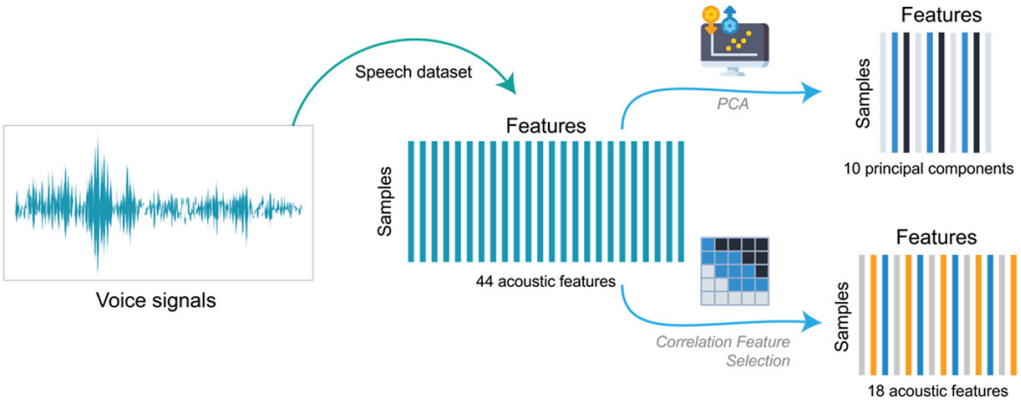

Fig. 2. Feature selection process

The process illustrated in Figure 2 outlines how the classification was applied to the processed voice data, incorporating both PCA and CFS to achieve predictive results. The CFS method searches for pairs of highly correlated measures similar to [29] and removes an arbitrary measure from each pair. The PCA method reduces the dimensionality of the dataset by identifying combinations of features that capture the most variance in the data. Additionally, PCA removes correlated measures by transforming the original features into a new set of uncorrelated variables, known as principal components. This process helps to streamline the feature space while retaining the most important information for classification tasks [15], [22].

2.6 Classification

Considering the studies presented in the literature [2], [10], [12], [15], [28], [30], we used a broad supervised ML modeling approach, implementing 19 algorithms to classify patients with PD. We used the default set of classification models available in the PyCaret library, encompassing a comprehensive range of widely used algorithms for classification problems. Additionally, we included models such as CatBoost, XGBoost, MLP, and RBF-SVM, which are extensively cited in the literature [2], [17]. This approach ensured a robust comparison and minimized potential biases in model selection by considering a diverse set of algorithms.

Our framework is comparable to that used in [16]. The algorithms initially submitted to the training data were the Ada boost classifier (ADA), CatBoost classifier (CATBOOST), decision tree classifier (DTC), dummy classifier (DUMMY), extra trees classifier (ETC), extreme gradient boosting (XGBOOST), Gaussian Process classifier (GPC), gradient boosting classifier (GBC), K neighbors classifier (KNN), light gradient boosting machine (LGHTGBM), linear discriminant analysis (LDA), logistic regression (LR), MLP classifier (MLP), Naïve Bayes (NB), quadratic discriminant analysis (QDA), Random Forest (RF) classifier, ridge classifier (RIDGE), support vector machine-linear kernel (SVM), and support vector machine-radial kernel (RBFSVM).

To mitigate the risk of overfitting, we employed a comprehensive approach that included cross-validation, hyperparameter tuning, and FS techniques. After training all models on the training dataset, we used k = 5 folds to verify the models' performance, as the concept used in [2], [10], [12], and [16]. This process involved dividing the training data into five subsets and training the model on four subsets while testing it on the remaining subset, rotating this process to ensure each subset was used as a testing set. This rigorous testing on different subsets prevented the models from becoming overly tailored to the training set. The performance of the algorithms was presented in terms of average accuracy and standard deviation. This process requires significant computational resources for hyperparameter optimization. Initially, we set aside 7.5% of the dataset for validation purposes. This step helps mitigate overfitting by providing an additional evaluation phase separate from training and testing. The remaining 222 records from the original 240 were then divided into 80% for training and 20% for testing to assess the models' generalization capabilities. However, it is necessary, as this process contributes to obtaining accurate results and reducing overfitting when the considered dataset is small [12], [31]. We integrated FS techniques like CFS and PCA to further enhance robustness and curb overfitting. These methods not only aid in fine-tuning model parameters but also contribute to creating more resilient models, aligning with established practices in predictive modeling [32].

Different metrics exist in medical diagnosis systems for measuring the performance of classification methods, such as accuracy, recall, precision, f-score, and Matthews

Correlation Coefficient (MCC) [7]. Accuracy is widely used in literature [2], [10], [12], [16], [28], [33] as a model performance metric. Based on the values derived from the confusion matrix, several key metrics can be calculated. The four possible outcomes are true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [10]. These outcomes allow us to compute various performance indicators: accuracy as shown in Equation 1; precision, which measures the proportion of cases identified as positive by the classifier, as defined in Equation 2; recall, indicating the percentage of correctly identified positive samples as detailed in Equation 3; f-score, which provides a combined measure of precision and recall, described in Equation 4; and the MCC, which is particularly useful for datasets with imbalanced class sizes, as outlined in Equation 5 [34].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F - Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \tag{4}$$

$$MCC = \frac{TN * TP - FN * FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

2.7 The area under ROC curve

To compare the efficiency of the models, we used accuracy as one of the metrics. For the model with the highest accuracy, we generated the receiver operating characteristic (ROC) curve, which is formed by plotting the FP rate (FPR) on the x-axis and the TP rate (TPR) on the y-axis [35]. The ROC curve is a graphical plot that illustrates the diagnostic trade-off between clinical recall/sensitivity and specificity for every possible cutoff in a combination of binary classification tests.

The area under the curve (AUC) estimates the area underneath the entire curve and is commonly used to assess the diagnostic performance in several biomedical applications [15], [36]. The expressions for TPR and FPR are defined as follows:

$$TPR = \frac{TP}{TP + FN} \tag{6}$$

$$FPR = \frac{FP}{TN + FP} \tag{7}$$

Where *TP*, *FN*, *FP,* and *TN* are used to calculate the rates [35].

2.8 Data preprocessing and implementation

Data preprocessing and implementation of ML algorithms were performed using the Python programming language version 3.10.12 and Visual Studio Code software version 1.88.0. Default hyperparameters were used for each model to replicate the results of future studies. A commonly used technique to evaluate the performance of

ML models from a dataset is to run a training process by testing multiple times and calculating the mean and standard deviation for the initially defined performance metric, which is the accuracy of this study. The functions used to configure, create, test, and optimize the model's performance multiple times were 'setup', 'compare_models,' 'create_model,' and 'tune_model,' from the PyCaret library, which performs cross-validation according to a previously defined number of partitions [2], [12], [16], [37].

Specifically, the 'setup' function was used to prepare the data for model training. This function offers a variety of parameters, enabling the creation of a comprehensive data preprocessing pipeline. Conversely, the 'compare_models' function simplifies the process of selecting the best classification model by training all available models and presenting the results in an accessible format. The 'create_model' function allows the creation of a specific model of interest with more flexibility. Lastly, the 'tune_model' function plays a crucial role in enhancing the performance of a classification model through hyperparameter tuning, a technique aimed at finding the optimal combination of hyperparameters. It adjusts the hyperparameters of a given estimator, producing a score grid with cross-validation scores for the best model [27], [38]. After the optimization process, the 'predict_model' function is applied to the models, producing, by default, predictions over the testing data that is used to determine how well the model generalizes to new data. Also, it is possible to change the parameter of this function to apply this to the validation data, considering 'predict_model(ml_model, data = validation_data)' that in this study is represented by (r_{val}, v_{val}).

3 RESULTS

3.1 Classification algorithms applied to all features

The results highlighted in Table 4 were obtained by implementing the algorithms with default hyperparameters using the PyCaret library. For each model, we assumed a value for the average accuracy in percentage, with train size = 0.80, fold strategy = stratifiedkfold, fold = 5, normalized data by z-score with normalized = True, remove_multicollinearity = False, and all the other hyperparameters from the 'setup' function by default. Additionally, the parameter 'session_id = 1245' was set to ensure reproducibility of results.

Table 4. Performance analysis of 19 models applied to training data considering all features (mean values from k-fold CV)

Model	Training Data Mean of Accuracy, Recall, Precision, F-score (%), AUC ([0,1]) and MCC ([–1,1])					
	Accuracy (%)	Recall (%)	Prec. (%)	F-score (%)	AUC	MCC
RBFSVM	84.14	0.8881	76.21	91.04	0.8264	0.6959
NB	83.56	0.8688	79.54	86.10	0.8253	0.6748
GPC	82.44	0.8135	79.61	84.74	0.8194	0.6527
ET	82.43	0.8983	78.37	85.36	0.8128	0.6559
CATBOOST	81.87	0.8888	76.14	86.42	0.8051	0.6477
RF	81.32	0.8548	78.37	83.96	0.8040	0.6377
LIGHTGBM	80.21	0.8638	75.10	84.52	0.7903	0.6142

(Continued)

Table 4. Performance analysis of 19 models applied to training data considering all features (mean values from k-fold CV) *(Continued)*

Model	Training Data Mean of Accuracy, Recall, Precision, F-score (%), AUC ([0,1]) and MCC ([–1,1])					
	Accuracy (%)	Recall (%)	Prec. (%)	F-score (%)	AUC	MCC
MLP	79.57	0.8786	80.72	79.96	0.8009	0.5948
KNN	79.10	0.8374	75.03	82.10	0.7822	0.5868
XGBOOST	79.08	0.8743	75.03	82.04	0.7806	0.5879
RIDGE	79.05	0,0000	75.03	82.97	0.7816	0.5911
ADA	78.49	0.8061	76.21	80.21	0.7796	0.5734
LR	78.43	0.8901	76.01	80.03	0.7793	0.5696
GBC	77.97	0.8429	72.81	81.30	0.7649	0.5661
LDA	77.33	0.8308	73.86	80.49	0.7657	0.5544
SVM	74.57	0,0000	69.41	78.39	0.7332	0.4989
QDA	70.67	0.7765	55.69	80.78	0.6486	0.4389
DT	69.46	0.6948	65.95	70.48	0.6785	0.3922
DUMMY	50.29	0.5000	0.00	0.00	0.0000	0.0000

In order to calculate the relevant metrics, the confusion matrix corresponding to each model was obtained using the algorithm, from which the expected and true outcomes can be obtained. Figure 3 represents the top three best-performing models on the test subset: NB, QDA, and GPC.

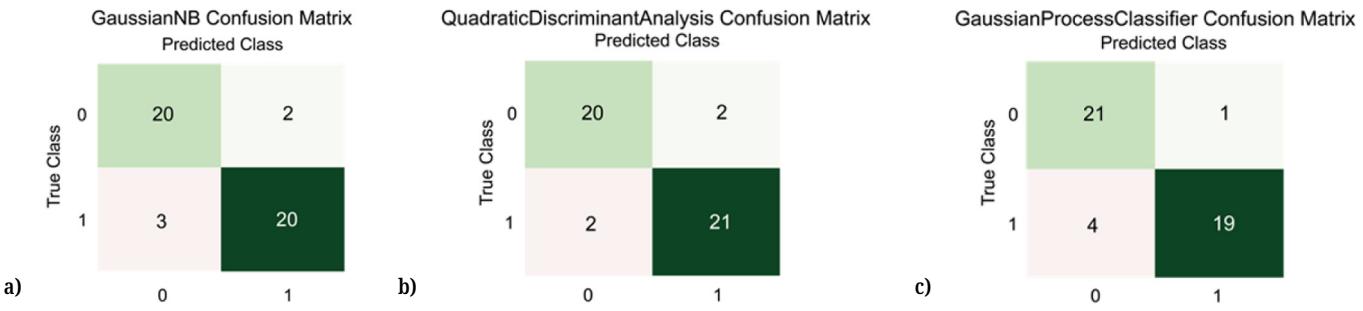

Fig. 3. Confusion matrices for the top three models applied to the testing dataset: (a) GaussianNB (NB), (b) Quadratic Discriminant Analysis (QDA), and (c) Gaussian Process Classifier (GPC)

Using the values presented in Figure 3 for the NB model on the testing data with the expressions for accuracy, recall, precision, F-score, and MCC defined in [10], [15], it was possible to calculate this value for the top six models. The hyperparameters of each model were optimized using the 'tune_model' function from the PyCaret library, which can improve the performance of a classification model by determining the optimal combination of its hyperparameters tuning [27], [38]. We compared the performance of each algorithm when subjected to test data on the accuracy metric. Table 5 presents accuracy, recall, precision, F-score, and MCC values for the best-performing models on the testing data, following the optimization of hyperparameters across all 19 models. After the optimization of the 19 ML models, we have the results presented in Table 5, considering the testing data.

Table 5. Performance analysis for the best six models applied to testing data for all features

Model	Testing Data Mean of Accuracy, Recall, Precision, F-score (%), AUC ([0,1]) and MCC ([–1,1])					
	Acc.	Recall	Prec.	F-score	AUC	MCC
QDA	91.11	91.30	91.30	91.30	0.9466	0.8221
NB	88.89	86.96	90.91	88.89	0.9200	0.7787
GPC	88.89	82.61	95.00	88.37	0.9130	0.7853
CATBOOST	88.89	82.61	95.00	88.37	0.9486	0.7853
LDA	86.67	78.30	94.74	85.71	0.9526	0.7461
MLP	86.67	78.26	94.74	85.71	0.9486	0.7461

The criterion for selecting the best-performing models was based on each model's performance on the training data, considering the mean accuracy value and, for tie-breaking among models with the same accuracy value, the F-score metric value was considered. Model optimization was conducted using the default parameters of the 'tune_model' function for all 19 ML algorithms in the PyCaret library.

3.2 Classification algorithms after using feature selection

Feature selection has proven to be a successful preprocessing tool for ML problems [39]. However, it is difficult to select from the growing number of available models. In this study, two FS methods were used, motivated by the results obtained in the literature [2], [7], [16], [28], [39], and the considerable increase in the accuracy value [10] after applying FS techniques. The CFS, defined in [39], was utilized to select highly correlated voice features employing a correlation coefficient threshold greater than or equal to 0.90 ($r \geq 0.90$). The correlation values of the removed features ranged from 0.91 to 0.99, confirming the high correlation. To illustrate this, Figure 4 provides a heat map showing the correlations among the 18 features that were retained after applying the CFS method. In total, 27 variables were removed, and the final set of features, including those selected and those eliminated, are detailed in Table 6, except for the target variable, Status.

Table 6. The features removed and the selected after the technique of CFS

Method	Features Removed	Features Selected
CFS	Jitter_rel, Jitter_abs, Jitter_RAP, Shim_loc, Shim_dB, Shim_APQ3, Shim_APQ5, HNR05, HNR15, HNR25, HNR35, MFCC0, MFCC1, MFCC4, MFCC6, MFCC7, MFCC8, MFCC9, MFCC10, MFCC11, MFCC12, Delta0, Delta3, Delta4, Delta5, Delta6, Delta7	Gender, Jitter_PPQ, Shim_APQ11, HNR38, RPDE, DFA, PPE, GNE, MFCC2, MFCC3, MFCC5, Delta1, Delta2, Delta8, Delta9, Delta10, Delta11, Delta12
Number	27	18

Principal component analysis was applied to the initial set of 46 variables, resulting in a new subset of variables by restricting the analysis to 10 components, and the PCA method utilized in the 'setup' function from the PyCaret library was

"pca_method = 'kernel.'" These components represent linear combinations of the original variables, capturing most of the data's variability while preserving essential characteristics.

Afterward, to select the most relevant features, they were applied to all models to measure the information gained regarding the class for CFS and PCA. The CFS technique was strategically employed on the original set of 46 features to formulate a concise yet powerful model. Subsequently, the multitude of acoustic features was distilled to a more manageable 18, which were subsequently evaluated using 19 distinct classification methods. The visual representation in Figure 4 depicts a heat map illustrating the correlations among these 18 features, comprising 17 acoustic features.

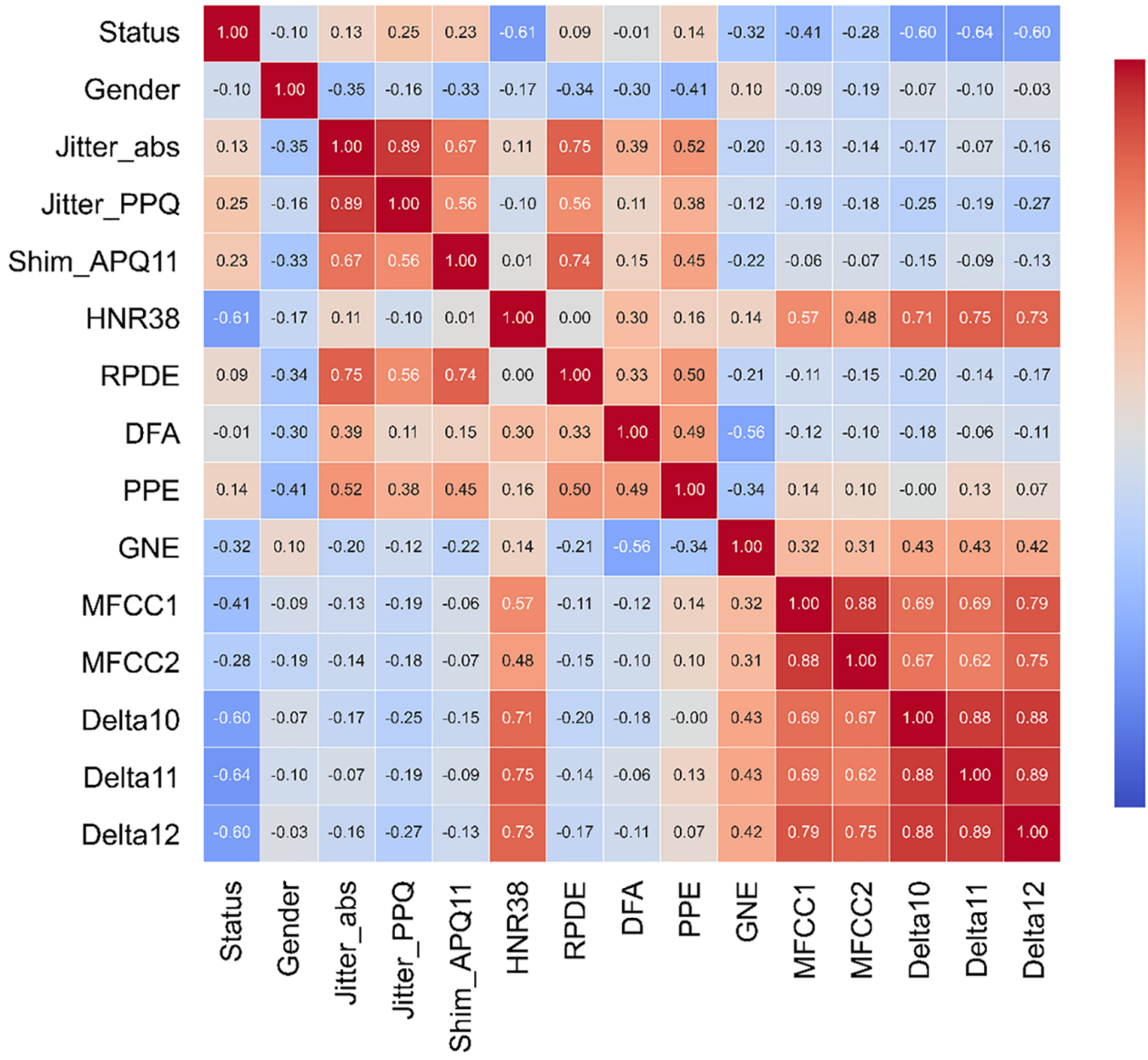

Fig. 4. Heat map showing the correlation among the 18 features after the CFS

The criterion for selecting the best-performing models presented in Table 7 is related to the performance of each model on the testing data, which were optimized

using the function 'tune_model,' as mentioned in this study, considering the mean value and standard deviation for the accuracy metric.

Table 7. Model performance analysis applied to training and testing data

Method	Model	Training Data		Testing Data
		Before FS Mean Accuracy (%)	After FS Mean Accuracy (%)	After the FS Process Mean Accuracy (%)
CFS	LDA	77.33	81.29	93.33
	GPC	82.44	80.75	91.11
	KNN	79.10	79.10	91.11
	SVM	74.57	76.79	91.11
	RIDGE	79.05	81.29	91.11
	MLP	79.57	75.67	88.89
PCA	NB	83.56	83.60	88.89
	RBFSVM	84.14	81.89	88.89
	ETC	82.43	82.44	88.89
	GPC	82.44	81.89	86.67
	LR	78.43	83.00	86.67
	KNN	79.10	84.13	86.67

The model performance analysis applied to the training and testing data, before and after the FS process, demonstrated that a great number of models achieved better performance after the FS process (CFS and PCA) considering the training data, except for RBFSVM and GPC using the PCA technique over the training data, where the mean value for accuracy was 84.14% before the FS and reduced to 81.89% and 82.44% that reduced to 81.89%, respectively. After the CFS method, the mean value was 82.44%, reduced to 80.75% for the GPC model, and for MLP, reduced from 79.57 to 75.67. The analysis of the top six models on the testing data revealed that all models performed well with an accuracy metric between 86.67% and 93.33%.

Figure 5 shows the feature importance plot for LDA and SVM algorithms after the CFS technique, which presents the 10 most important features for each model. Figure 6 shows the ROC curve for the models GPC, LDA, NB, and SVM, which obtained relevant values for the AUC. The FS process aimed to streamline the initial set of 44 acoustic features to develop a concise and efficient model, employing the CFS and PCA techniques.

Subsequently, the application of the PCA technique led to a reduction in the number of acoustic features to 10 principal components, and the CFS technique reduced the number of 44 to 17 acoustic features. These selected features were then utilized in 19 distinct classification methods.

The Feature Importance plot, after the CFS technique, assigns a score to each feature, where higher scores signify increased relevance to the output data variable [7], [16].

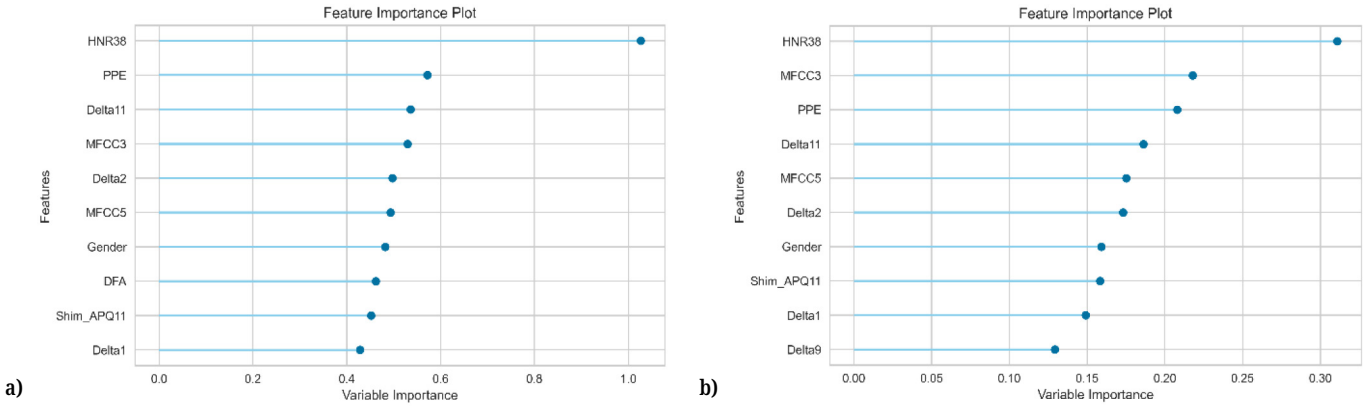

Fig. 5. The ten features of more importance for: a) LDA, b) SVM algorithms after the application of the CFS technique

Figure 5, panels a) and b), display the feature importance profiles for models LDA and SVM applied to the same subset of 18 features derived from an initial set of 45 via the CFS technique. Notably, within this subset, both models, despite their inherent algorithmic differences, share common variables among the five most important. Considering the LDA, the most impactful are 'HNR38', 'PPE,' 'Delta11,' 'MFCC3,' and 'Delta2,' while for SVM, they are 'HNR38,' 'MFCC3,' 'PPE,' 'DELTA11,' and 'MFCC5.'

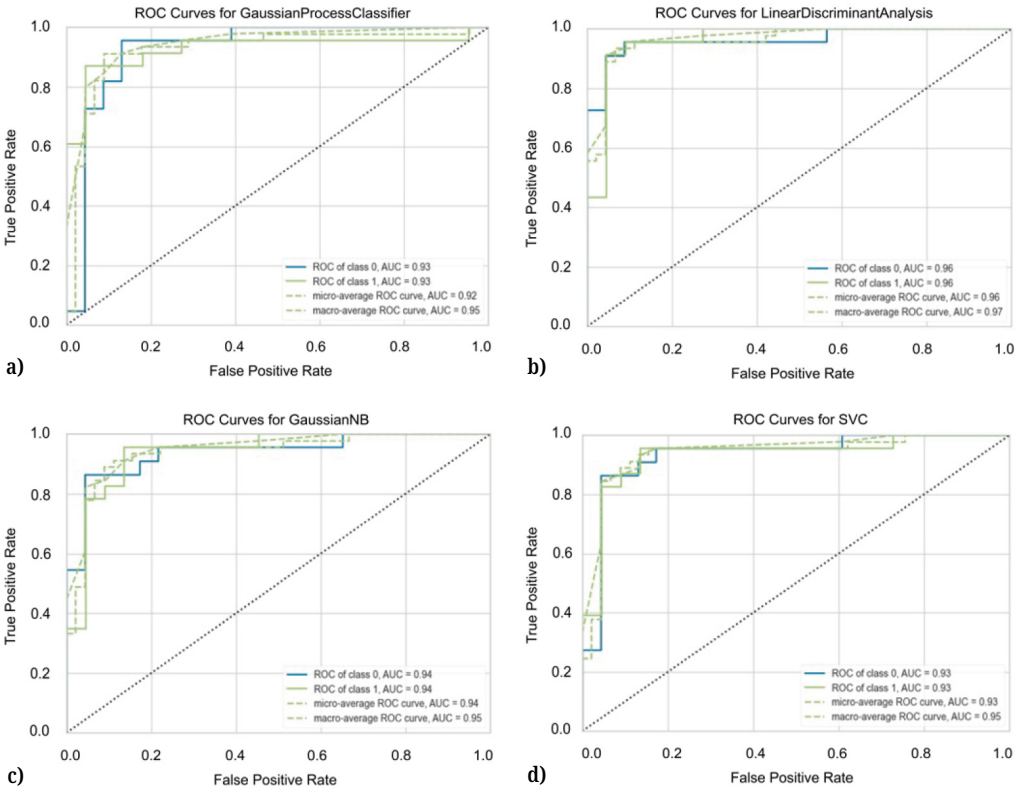

Fig. 6. ROC curve generated by a) GPC and b) LDA after the application of the technique of CFS, c) NB and d) SVM-Linear after the application of the PCA

After applying the CFS method, the estimated AUC values of the GPC model were 0.93, and the LDA's were 0.96 for the ROC of classes 0 and 1. Considering the NB after using the PCA method, the AUC values were 0.94; the SVM was 0.93 for the ROC of classes 0 and 1, thereby representing a promising result since the maximum value is 1.

Table 8 summarizes the results of classification experiments using different methods and models with distinct subsets of features. Each FS or dimensionality reduction method was applied to the training, testing, and validation datasets, and the models were evaluated based on their performance metrics, which in this study was considered the accuracy.

Table 8. Model performance analysis applied to training, testing, and validation data with the Accuracy given in %

Stage	Model	Subset for		
		Training Mean Accuracy (%)	Testing Mean Accuracy (%)	Validation Mean Accuracy (%)
All features	GPC	82.44	88.89	83.33
	XGBOOST	79.08	86.67	72.22
	KNN	79.10	82.22	77.78
Correlation Feature Selection	GPC	82.44	91.11	83.33
	ETC	82.43	88.89	77.78
	KNN	79.10	88.89	77.78
Principal Component Analysis	DTC	69.46	80.00	83.33
	ETC	82.43	88.89	72.22
	KNN	79.10	86.67	66.67

The models include GPC, KNN, DTC, XGBOOST, and ETC; their results provide a comparative view of the models' performance across different feature configurations. Table 9 presents the outcomes from ten studies that engaged the identical dataset for data classification purposes, adopting a methodology akin to ours. Even though each of these studies deployed distinct classification techniques, strategies for optimizing hyperparameters, FS techniques, and proportions for training, testing, and validation data, they uniformly analyzed the same volume of records, totaling 240.

Table 9. Comparison of the accuracy between the proposed method in this work and previous works that apply classification to the same dataset

References	FS Technique	Number of Features	Train Test Validation	Classify Method	Acc. (%)
Ghaheri et al. (2023) [17]	Pearson Correlation	32	75/25	XGBoost	85.42
Özseven et al. (2023) [18]	–	–	80/20	BA-SVM	86.70
				ICPSO-SVM	88.80
				IGA-SVM	88.30
				HSA-SVM	88.30
Islam et al. (2022) [19]	–	44	70/15/15	FFNN	85.00
Bao et al. (2022) [15]	PCA FS	44	90/10	KNN	80.60
				SVM	82.50
				SSCL	83.80

(Continued)

Table 9. Comparison of the accuracy between the proposed method in this work and previous works that apply classification to the same dataset *(Continued)*

References	FS Technique	Number of Features	Train Test Validation	Classify Method	Acc. (%)
Saeed et al. (2022) [20]	PCA IG	45 to 20	–	KNN	88.33
Fahim et al. (2021) [21]	CFS, ANOVA, CSFS, MIFS, RFE	–	–	GNB	82.50
				LR	83.11
Mittal et al. (2021) [22]	PCA	45	–	wkNN	90.30
Yaman et al. (2020) [23]	Statistical for Feature Increasing	45 to 177	–	KNN	91.23
	ReliefF algorithm	177 to 66		SVM	91.25
Karabayir et al. (2020) [2]	FI and Incremental Feature Addition	45 to 265	–	LGB	84.10
				RF	81.80
		265 to 15		XGB	81.80
				LR	77.10
Bielby et al. (2020) [24]	–	45	70/30	LRNN	96.00
Yasar et al. (2019) [25]	–	45	80/15/5	ANN	94.93
Perez et al. (2016) [26]	FS using Statistical analysis	45 to 27	90/10	RBF	85.25
				Linear	84.29
				MLP	81.51
				Quadratic	79.50
Our proposed work	All features	45	74/18/8	QDA	91.11
				GPC	88.89
	CFS	45 to 18		LDA	93.33
				GPC	91.11
	PCA	45 to 10		NB	88.89
				RBFSVM	88.89

In comparison with prior studies employing the same dataset, as shown in Table 9, a variety of approaches were observed. While some studies [17], [20], [21], [22], [23], [26] utilized techniques for FS, such as Pearson correlation, PCA, ANOVA, CFS, statistical analyses, and others, we have some studies that did not adopt any specific strategy in this regard. Moreover, it is noteworthy that only [19] and [25] works incorporated validation sets into their analyses, akin to our study. Regarding performance, several studies achieved significant accuracy ranging from 85% to 96%. Particularly, our study attained an accuracy of 91.11% considering all features and 93.33% after the CFS technique, surpassing many prior studies.

4 DISCUSSION

The architecture was divided into two main methods: all feature analysis and FS methods, and PD prediction in all steps described in Figure 1, starting with pre-processing until the final verification of the performance of the classification models, considering both testing and validation subsets. Based on the results obtained,

we classified patients with PD by analyzing voice recordings using ML as a noninvasive diagnostic tool that uses acoustic resources extracted from speech recordings according to the procedure described in [11].

The first analysis was obtained after applying nineteen models to classify all features, initially resulting in an accuracy of 84.14% for RBFSVM, 83.56% for NB, 82.44% for GPC, and 82.43% for ETC, considering the training data. After applying the 'tune_model' function for all 19 models, we obtained better accuracy values for the testing data of 91.11% for QDA, 88.89% for NB, CATBOOST, and GPC, and 86.67% for LDA and MLP. Our findings indicated that all models achieved excellent performance using the F-score metric, which represents the balance between precision and recall, with values ranging from 85.71% to 91.30%.

Theoretical insights, as explored by [40] and [41], propose that the expansion of the feature set size poses challenges to reliable classification. This is attributed to the reduced coverage of the feature space by measures from a fixed number of subjects. Therefore, addressing these challenges requires the implementation of FS strategies. The recommendation is to trim the feature set down to a minimal size that preserves the optimal information necessary for effective classification, as advocated by [32], [41]. Our study also used two FS methods, parallel to the analysis considering all features: CFS and PCA, using the features extracted from the speech signals.

Based on the results presented in Table 7, it can be inferred that the FS has a positive effect on the performance of the classifiers, as shown in [2], [10], and [16]. The two FS methods exhibited different performances for the 19 classifiers considered in this study. Importantly, our findings demonstrate that FS techniques can be efficient data preprocessing tools to reduce data dimensionality and identify the most significant risk factors for patient classification in PD. The variables "HNR38," "PPE," "MFCC3," "MFCC5," "Delta2," and "Delta11" were found to be particularly relevant for PD determination by CFS, as demonstrated in Figure 5. Notably, "HNR38" and "PPE" ranked within the top three variables for both LDA and SVM models, highlighting their importance in PD classification. These results underscore the significant impact of FS techniques in enhancing the classification performance of LDA and SVM models, contributing to a deeper understanding of the acoustic features crucial for PD diagnosis. Little (2009) [41] introduced the pitch period entropy (PPE) as a new measure of dysphonia, robust to many uncontrollable confounding effects, including noisy acoustic environments and normal, healthy variations in voice frequency. Furthermore, it is observed in his study [41] that the combination of HNR, RPDE, DFA, and PPE achieves the best overall classification performance. Our findings, complementing Little's work, confirm the importance of PPE and HNR in distinguishing PD patients, thereby emphasizing the robustness and effectiveness of these acoustic features in PD diagnosis.

The non-linear group features (RPDE, DFA, PPE, GNE), defined in Table 2, may reflect the complexity and irregularity of vocal production in patients with hypokinetic dysarthria. Disordered sustained vowels present a variety of phenomena, from nearly periodic vibrations to highly complex and aperiodic vibrations, which is precisely what the analyzed features seek to quantify [41], [42]. Regarding the MFCCs and Delta Coefficients groups, also defined in Table 2, they can estimate the voice's spectral variations associated with changes in vocal quality and articulation [11], being variables with significant weight for the models of better performance, as observed in Figure 5, items a) and b). Lastly, the Harmonic-to-Noise ratio group indicate, that the voice may have an increase in noise, reducing the HNR, indicating a more unstable and unpredictable voice [11]. As supposed, a great number of models showed good accuracy when applied to training data after the FS stage presented

in Figure 2, compared to the process considering all features, with emphasis on LDA, which improved its accuracy by 5.12% after the CFS method, and LR and KNN, which improved their accuracy by 5.83% and 6.36%, respectively, after the PCA method. All six top models had an accuracy between 88.89% and 93.33% for the testing data considering the CFS technique presented and an accuracy between 86.67% and 88.89% considering the PCA technique for feature selection.

Our experimental AUC results are comparable to those of other notable studies in the literature [15], [28]. In this study, GPC, LDA, NB, SVM, ETC, and RFC produced good results for the AUC with a value greater than 0.9200, and MLP produced better overall AUC results for the different FS methods considered, with 0.9756 after the CFS and 0.9545 after the PCA. To demonstrate this, the ROC curves for GPC, LDA after CFS, NB, and SVM after PCA are shown in Figure 6.

The positive outcomes depicted in Figure 6 and detailed in Table 7 highlight the congruence between the performance of the RBFSVM algorithm and the objectives and anticipated results of our study as recommended by [40], [41]. Through the utilization of Gaussian radial basis kernel functions, recognized for their flexibility in creating smooth, curved decision boundaries [41], our findings indicate that optimal decision boundaries for distinguishing between healthy and PD individuals may not adhere to simple lines or hyperplanes as the result compared to the linear kernel SVM algorithm. The analysis of model performance on training, testing, and validation data, as depicted in Table 8, reveals intriguing insights. Despite not consistently leading in training and testing accuracy across all stages of the analyses considered in this study (firstly using all features and then using the two FS methods), as observed in Tables 5 and 7, the GPC consistently demonstrated superior performance on validation data. Notably, GPC achieved the highest accuracy on validation data in both: All Features and CFS stages, showcasing its robustness and effectiveness in PD classification. These results underscore the importance of considering validation data to assess model generalization and reliability across different FS techniques.

Observing the results presented in Table 9, our study brings a comprehensive analysis of the classification of PD using ML techniques, leveraging a dataset widely utilized in previous studies. We explored two FS techniques besides the analysis of all features, CFS and PCA, to identify the most relevant features for PD classification. Our proposed methodology achieved promising results, with accuracies of 91.11% considering all features, 93.33% with CFS, and 88.89% with PCA on the test dataset. Notably, our approach incorporated an innovative dataset division into training, testing, and validation subsets, enabling a robust evaluation of model performance. Additionally, our study considered analyses for different numbers of folds, as observed in [2], ranging from 4 to 10 folds, with the 5-fold strategy yielding the best result. Moreover, proportions of 5%, 7.5%, and 10% on the initial dataset of 240 records were explored, with the 7.5% proportion yielding the most significant results considering the validation data as your final focus. Our study contributes to the literature by comparing our results with those of previous studies, highlighting the superiority of our proposed method in terms of accuracy. Overall, our findings underscore the efficacy of ML models in PD classification and emphasize the importance of thoughtful FS and dataset partitioning in achieving accurate and reliable results.

The GPC outperformed other models during testing on the validation data, despite some models showing better results during training. This could be because the GPC is a non-parametric model that adjusts its complexity based on the data, potentially leading to better generalizations on the unseen. Moreover, the probabilistic nature of the GPC allows it to effectively handle uncertainty and variations in the data,

which can lead to better performance on validation datasets. Models such as SVM, XGBoost, NB, and RF have also shown good performance in the studies presented in Table 9, where SVM [15], [18], [23], [26], XGBoost [17], NB in [21], KNN [15], [20], [23], and RF [2] demonstrated notable results.

Even though our study did not use sensitive patient data and the data collection was approved by the Bioethical Committee of the University of Extremadura, as cited in Section 2.1, it is crucial to address the ethical implications of applying ML in healthcare. Protecting patient privacy and data confidentiality, as well as respecting patient autonomy, are essential. Informed consent is necessary to ensure that patients understand how their data will be used. Additionally, transparency in how algorithms make decisions is vital for maintaining trust and ensuring fairness. By considering these ethical issues, we can support the responsible use of ML in clinical settings, enhancing patient care while safeguarding patient rights.

4.1 Limitation of the study

Our study has some limitations that merit discussion. Firstly, the small sample size may restrict the generalizability of our findings and limit the robustness of our conclusions. Additionally, the composition of our dataset, which solely includes subjects from the control group and those diagnosed with PD, may present challenges in distinguishing between PD and other conditions affecting voice characteristics. Unfortunately, the dataset used in this study lacked clinical variables, limiting our ability to evaluate disease progression or the response to pharmacologic treatment in PD. We acknowledge the importance of these factors and recognize their significance in clinical practice. Therefore, future studies should aim to incorporate comprehensive datasets encompassing clinical markers to provide a more thorough evaluation of disease progression and treatment outcomes. Relying solely on speech for an accurate diagnosis of PD is insufficient; it is essential to consider other symptoms to improve diagnostic accuracy.

We employed cross-validation and separated a validation subset to monitor and mitigate this risk, but further validation with larger and more diverse datasets is recommended. Additionally, we did not account for potential confounding variables, such as age, sex, medications, and comorbid conditions, which could influence vocal disorders in PD patients. Addressing these confounders in future research will be crucial to improving the precision and generalizability of our findings. Despite these limitations, our study contributes valuable insights into the potential of ML algorithms in assisting clinicians with the diagnosis of PD using voice analysis. Moreover, the framework established in this study in Figure 1 can serve as a basis for future investigations, emphasizing the importance of transparency and reproducibility in algorithmic development for clinical applications.

5 CONCLUSION

This study utilized various ML models to identify patients with PD through a non-invasive speech test, aiming to discriminate between PD patients and healthy controls based on acoustic features extracted from voice recordings and to identify the most important combination of voice features for early PD classification. Employing two FS algorithms from the SciKit-Learn and Pycaret libraries, we analyzed 19 models using voice signal features from both PD patients and healthy individuals for early

PD diagnosis. Our primary objective was to enhance model performance, accuracy, and computational efficiency in the classification task.

Classification accuracy was assessed considering all variables. Following the FS process and identification of the top six classification models for each FS technique, we achieved outstanding performance using LDA, GPC, KNN, SVM with Radial Kernel, and Multi-layer Perceptron with CFS. Additionally, NB, SVM with Radial Kernel, ETC, GPC, LR, and KNN showed excellent results after applying PCA. Notably, the LDA model achieved the highest testing classification accuracy of 93.35% with CFS, while NB achieved 88.89% accuracy with PCA. GPC yielded the highest validation classification accuracy of 83.33% over the validation data, considering all features, and after the CFS technique.

The remarkable performance of our models, particularly GPC, across subsets (training, testing, and validation) and after FS techniques not only aligns with but also surpasses that of many existing studies using the same dataset for PD diagnosis through voice analysis, as observed in Table 9. This underscores the significant progress made in applying ML to enhance the precision and efficiency of noninvasive PD diagnostics. Although our study demonstrates a significant step forward in showcasing the potential of ML in voice analysis for PD detection using 19 different ML models, applying these findings in real-world healthcare settings presents several challenges. These challenges include integrating ML models into existing healthcare systems and training clinicians to effectively use and interpret these models. While our study marks a significant step in demonstrating ML's potential in voice analysis for PD detection, further external validation and in-depth studies are necessary to confirm these results with recordings of normal conversations. Such investigations are crucial to affirming the practical viability of these ML models in clinical settings, potentially revolutionizing the early detection and ongoing monitoring of PD. By integrating various data types and advancing the practical application of these models, we aim to improve the reliability and accessibility of PD diagnostics, ultimately benefiting patient care.

6 ACKNOWLEDGMENTS

The authors of this work are grateful to the National Council for Scientific and Technological Development (CNPq); the Coordination of the Improvement for Higher Level Personnel (CAPES); the Ministry of Science, Technology, Innovation, and Communications (MCTIC); the Research Foundation of the State of Minas Gerais (FAPEMIG); the Federal Institute of Education, Science, and Technology Goiano (IF Goiano) for the financial and structural support to conduct this study; and the UCI ML repository for making available the database used in this study. A. A. Pereira is a research productivity fellow at CNPq, Brazil (309525/2021-7).

7 REFERENCES

[1] S. Cerri, L. Mus, and F. Blandini, "Parkinson's disease in women and men: What's the difference?" *J. Parkinsons Dis.*, vol. 9, no. 3, pp. 501–515, 2019. https://doi.org/10.3233/JPD-191683

[2] I. Karabayir, S. M. Goldman, S. Pappu, and O. Akbilgic, "Gradient boosting for Parkinson's disease diagnosis from voice recordings," *BMC Med. Inform. Decis. Mak.*, vol. 20, p. 228, 2020. https://doi.org/10.1186/s12911-020-01250-7

[3] J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008. https://doi.org/10.1136/jnnp.2007.131045

[4] M. T. Hayes, "Parkinson's disease and Parkinsonism," *Am. J. Med.*, vol. 132, no. 7, pp. 802–807, 2019. https://doi.org/10.1016/j.amjmed.2019.03.001

[5] A. Joy, S. Menon, N. M. Thomas, M. Christy, A. D. Menon, and A. John, "Pharmacophore modelling and molecular dynamics simulation to identify novel molecules targeting catechol-O-methyltransferase and dopamine D3 receptor to combat Parkinson's disease," *Polymer Bulletin*, vol. 81, pp. 7893–7917, 2024. https://doi.org/10.1007/s00289-023-05087-8

[6] J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 79, pp. 368–376, 2008. https://doi.org/10.1136/jnnp.2007.131045

[7] D. Jain and V. Singh, "Feature selection and classification systems for chronic disease prediction: A review," *Egyptian Informatics Journal*, vol. 19, no. 3, pp. 179–189, 2018. https://doi.org/10.1016/j.eij.2018.03.002

[8] E. Ray Dorsey *et al.*, "Global, regional, and national burden of Parkinson's disease, 1990–2016: A systematic analysis for the global burden of disease study 2016," *Lancet Neurol.*, vol. 17, no. 11, pp. 939–953, 2018. https://doi.org/10.1016/S1474-4422(18)30295-3

[9] J. Massano and K. P. Bhatia, "Clinical approach to Parkinson's disease: Features, diagnosis, and principles of management," *Cold. Spring Harb. Perspect. Med.*, vol. 2, p. a008870, 2012. https://doi.org/10.1101/cshperspect.a008870

[10] Z. Karapinar Senturk, "Early diagnosis of Parkinson's disease using machine learning algorithms," *Med. Hypotheses*, vol. 138, p. 109603, 2020. https://doi.org/10.1016/j.mehy.2020.109603

[11] L. Naranjo, C. J. Pérez, Y. Campos-Roca, and J. Martín, "Addressing voice recording replications for Parkinson's disease detection," *Expert. Syst. Appl.*, vol. 46, pp. 286–292, 2016. https://doi.org/10.1016/j.eswa.2015.10.034

[12] I. Tougui, A. Jilbab, and J. El Mhamdi, "Machine learning smart system for Parkinson disease classification using the voice as a biomarker," *Healthc. Inform. Res.*, vol. 28, no. 3, pp. 210–221, 2022. https://doi.org/10.4258/hir.2022.28.3.210

[13] C. O. Sakar *et al.*, "A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform," *Appl. Soft. Comput.*, vol. 74, pp. 255–263, 2019. https://doi.org/10.1016/j.asoc.2018.10.022

[14] L. Naranjo, C. J. Pérez, J. Martín, and Y. Campos-Roca, "A two-stage variable selection and classification approach for Parkinson's disease detection by using voice recording replications," *Comput. Methods Programs Biomed.*, vol. 142, pp. 147–156, 2017. https://doi.org/10.1016/j.cmpb.2017.02.019

[15] G. Bao, M. Lin, X. Sang, Y. Hou, Y. Liu, and Y. Wu, "Classification of dysphonic voices in Parkinson's disease with semi-supervised competitive learning algorithm," *Biosensors (Basel)*, vol. 12, no. 7, p. 502, 2022. https://doi.org/10.3390/bios12070502

[16] R. Z. U. Rehman, S. Del Din, Y. Guan, A. J. Yarnall, J. Q. Shi, and L. Rochester, "Selecting clinically relevant Gait characteristics for classification of early Parkinson's disease: A comprehensive machine learning approach," *Sci. Rep.*, vol. 9, p. 17269, 2019. https://doi.org/10.1038/s41598-019-53656-7

[17] P. Ghaheri, H. Nasiri, A. Shateri, and A. Homafar, "Diagnosis of Parkinson's disease based on voice signals using SHAP and hard voting ensemble method," *Comput. Methods Biomech. Biomed. Engin.*, vol. 27, no. 13, pp. 1858–1874, 2023. https://doi.org/10.1080/10255842.2023.2263125

[18] T. Özseven and Z. Ş. Özkorucu, "Optimization of support vector machines for prediction of Parkinson's disease," *Measurement Science Review*, vol. 23, no. 1, pp. 1–10, 2023. https://doi.org/10.2478/msr-2023-0001

[19] R. Islam, E. Abdel-Raheem, and M. Tarique, "Voiced features and artificial neural network to diagnose Parkinson's disease patients," in *2022 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, 2022, pp. 132–136. https://doi.org/10.1109/ICECTA57148.2022.9990334

[20] F. Saeed *et al.*, "Enhancing Parkinson's disease prediction using machine learning and feature selection methods," *Computers, Materials & Continua*, vol. 71, no. 3, pp. 5639–5658, 2022. https://doi.org/10.32604/cmc.2022.023124

[21] M. I. Fahim, S. Islam, S. T. Noor, Md. J. Hossain, and Md. S. Setu, "Machine learning model to analyze telemonitoring dyphosia factors of Parkinson's disease," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 8, 2021. https://doi.org/10.14569/IJACSA.2021.0120890

[22] V. Mittal and R. K. Sharma, "Machine learning approach for classification of Parkinson disease using acoustic features," *J. Reliab. Intell. Environ.*, vol. 7, no. 3, pp. 233–239, 2021. https://doi.org/10.1007/s40860-021-00141-6

[23] O. Yaman, F. Ertam, and T. Tuncer, "Automated Parkinson's disease recognition based on statistical pooling method using acoustic features," *Med. Hypotheses*, vol. 135, p. 109483, 2020. https://doi.org/10.1016/j.mehy.2019.109483

[24] J. Bielby, S. Kuhn, S. Colreavy-Donnelly, F. Caraffini, S. O'Connor, and Z. A. Anastassi, "Identifying Parkinson's disease through the classification of audio recording data," in *2020 IEEE Congress on Evolutionary Computation (CEC)*, 2020, pp. 1–7. https://doi.org/10.1109/CEC48606.2020.9185915

[25] A. Yasar, I. Saritas, M. A. Sahman, and A. C. Cinar, "Classification of Parkinson disease data with artificial neural networks," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 675, no. 1, p. 012031, 2019. https://doi.org/10.1088/1757-899X/675/1/012031

[26] C. Perez, Y. C. Roca, L. Naranjo, and J. Martin, "Diagnosis and tracking of Parkinson's disease by using automatically extracted acoustic features," *J. Alzheimers Dis. Parkinsonism*, vol. 6, 2016. https://doi.org/10.4172/2161-0460.1000260

[27] PyCaret, "Low-code Machine Learning," 2023. [Online]. Available: https://www.pycaret.org

[28] A. Suppa *et al.*, "Voice in Parkinson's disease: A machine learning study," *Front. Neurol.*, vol. 13, 2022. https://doi.org/10.3389/fneur.2022.831428

[29] S. Yang *et al.*, "Effective dysphonia detection using feature dimension reduction and Kernel density estimation for patients with Parkinson's disease," *PLoS ONE*, vol. 9, no. 2, p. e88825, 2014. https://doi.org/10.1371/journal.pone.0088825

[30] A. Saeed, J. Baber, M. Z. Abbas, A. Sajid, H. Razzaq, and A. A. Khan, "Improving the imbalanced data accuracy using CNN and ReLU," *IETI Transactions on Data Analysis and Forecasting (iTDAF)*, vol. 2, no. 3, pp. 50–58, 2024. https://doi.org/10.3991/itdaf.v2i3.51013

[31] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLoS ONE*, vol. 14, no. 11, p. e0224365, 2019. https://doi.org/10.1371/journal.pone.0224365

[32] P. Bruce and A. Bruce, *Practical Statistics for Data Scientists: 50 Essential Concepts* (1st Ed.). Sebastopol, CA: O'Reilly Media, 2017.

[33] J. S. Almeida *et al.*, "Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques," *Pattern Recognition Letters*, vol. 125, pp. 55–62, 2019. https://doi.org/10.1016/j.patrec.2019.04.005

[34] T. A. Medeiros, R. G. Saraiva Junior, G. De Souza E Cassia, F. A. De Oliveira Nascimento, and J. L. A. De Carvalho, "Classification of 1p/19q status in low-grade gliomas: Experiments with radiomic features and ensemble-based machine learning methods," *Brazilian Archives of Biology and Technology*, vol. 66, 2023. https://doi.org/10.1590/1678-4324-2023230002

[35] R. C. Prati, G. E. de A. P. A. Batista, and M. C. Monard, "Curvas ROC para avaliação de classificadores," *IEEE Latin America Transactions*, vol. 6, no. 2, pp. 215–222, 2008.

[36] E. M. Smaili, M. Daoudi, I. Oumaira, S. Azzouzi, and M. E. H. Charaf, "Towards an adaptive learning model using optimal learning paths to prevent MOOC dropout," *International Journal of Engineering Pedagogy (iJEP)*, vol. 13, no. 7, pp. 128–144, 2023. https://doi.org/10.3991/ijep.v13i7.40075

[37] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition. New York, NY: Springer, 2009. https://doi.org/10.1007/978-0-387-84858-7

[38] G. Tolios, *Simplifying Machine Learning with PyCaret a Low-code Approach for Beginners and Experts!* 1st ed., vol. 1, Victoria, British Columbia, Canada: Leanpub Book, 2022.

[39] B. Remeseiro and V. Bolon-Canedo, "A review of feature selection methods in medical applications," *Comput. Biol. Med.*, vol. 112, p. 103375, 2019. https://doi.org/10.1016/j.compbiomed.2019.103375

[40] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY: Springer Nature, 2009. https://doi.org/10.1007/978-0-387-84858-7

[41] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Trans Biomed Eng*, vol. 56, no. 4, pp. 1015–1022, 2009. https://doi.org/10.1109/TBME.2008.2005954

[42] M. Little, "Parkinsons," UCI Machine Learning Repository, 2007. [Online]. Available: https://doi.org/10.24432/C59C74.

8 AUTHORS

Daniel Hilário da Silva is a student of the Postgraduate Program in Biomedical Engineering, at the Faculty of Electrical Engineering, Federal University of Uberlândia, and is working as a Lecturer at Instituto Federal de Educação, Ciência e Tecnologia Goiano – Campus Cristalina, Goiás, Brazil (E-mail: daniel.hilario@ifgoiano.edu.br).

Caio Tonus Ribeiro is a student of the Postgraduate Program in Biomedical Engineering, at the Faculty of Electrical Engineering, Federal University of Uberlândia, Minas Gerais, Brazil (E-mail: caiot@ufu.br).

Leandro Rodrigues da Silva Souza is a student of the Postgraduate Program in Biomedical Engineering, at the Faculty of Electrical Engineering, Federal University of Uberlândia, and is working as a Lecturer at Instituto Federal de Educação, Ciência e Tecnologia Goiano – Campus Rio Verde, Goiás, Brazil (E-mail: leandro.souza@ifgoiano.edu.br).

José Renato Munari Nardo is a student of the Postgraduate Program in Biomedical Engineering, at the Faculty of Electrical Engineering, Federal University of Uberlândia, Minas Gerais, Brazil (E-mail: jose.munari@ufu.br).

Adriano Alves Pereira is a Professor in the Postgraduate Program in Biomedical Engineering, at the Faculty of Electrical Engineering, Federal University of Uberlândia, Minas Gerais, Brazil (E-mail: adriano.pereira@ufu.br).