Biomedical Signal Processing and Control 66 (2021) 102415

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Analysis of voice as an assisting tool for detection of Parkinson's disease and its subsequent clinical interpretation

Gabriel Solana-Lavalle, Roberto Rosas-Romero *

Department of Electrical and Computer Engineering, Universidad de las Americas Puebla, Ex Hacienda Sta. Catarina Mártir S/N, San Andrés Cholula, Puebla, C.P. 72810, Mexico

ARTICLE INFO

Keywords:
Parkinson's disease detection
Voice analysis
Feature subset selection
Assisting tool for diagnosis

ABSTRACT

In this work, a voice-based analysis is conducted with the contribution of providing physicians with a decision tool along with framing information to help them see functional differences and understand why the detection method suspects PD. The voice-based detection method consists in applying *feature subset selection* and four different classifiers to voice recordings from five datasets (gender-based, balanced and unbalanced) derived from the largest public dataset for voice-based PD detection. One of the contributions is an improvement over previous works on voice-based PD detection over the same dataset, in terms of performance and complexity. The detection performance is characterized by 95.9% of *accuracy*, 98.35% of *sensitivity*, 91.06% of *specificity*, and 95.6% of *precision* in women; and 94.36% of *accuracy*, 100% of *sensitivity*, 97.1% of *specificity*, and 96.83% of *precision* in men. The number of features, fed to classifiers, ranges from 6 to 20. This work shows that different factors are associated with PD detection according to gender: high-frequency voice content is the most significant functional information to assist PD detection in women, while low-frequency content assists PD detection in men better. It is shown that a comparison of the variability of the most important features between patients with PD and controls can be used as contextual information by a physician to have a better interpretation of the classification.

1. Introduction

Parkinson's disease (PD) is the second most common chronic neurodegenerative disease after Alzheimer's disease. It affects 1% of the population over age 60, and it is estimated that the number of persons with Parkinson will double from 2005 to 2030 [1,2]. The diagnosis of PD is one of the most challenging tasks in neurology since PD is mostly diagnosed at an advanced stage [3].

According to Braak et al. [4], there are six neuropathologic stages during the evolution of the disease. Olfactory and vocal disorders appear at the first two stages. During the third and fourth stages, motor symptoms become more apparent, which facilitates the diagnosis of PD by an expert. Large areas of the brain are severely affected during the fifth and sixth stages. Some of the vocal impairments, in PD patients, are defective use of voice (dysphonia) [5], reduced volume (hypophonia), reduced pitch range (monotone) and difficulty with the articulation of sounds (dysarthria) [6].

Although PD is not cured, the quality of life of PD patients might be improved if prompt treatment is received as a consequence of an early diagnosis [7]. Vocal disorders are prodromal symptoms that manifest in 90% of the PD patients at early stages [6]. Thus, analysis of vocal features has been used for PD detection at early stages [3,7–11]. Furthermore, the vocal analysis method is inexpensive and easy to do early on.

The advantages of applying pattern recognition on voice recordings, as an assisting tool for PD diagnosis, are: (1) vocal disorders start in the first stages of the disease [12], (2) voice is easily recorded, (3) there is a wide diversity of signal processing techniques for extraction of voice features [13,14], (4) the classification of voice features with machine learning algorithms provides assistance during the early diagnosis of the disease [14,13], and (5) PD detection at early stages and the corresponding health cares might improve the quality of life of the patients [15].

According to the reported works on voice-based PD detection, the most commonly used group of vocal features has been the *baseline features* that include *jitter*, *shimmer*, *fundamental frequency parameters*, *harmonicity parameters*, *recurrence period density entropy* (RPDE), *detrended fluctuation analysis* (DFA) and *pitch period entropy* (PPE) [13]. Another comprehensive analysis of features, for PD detection, is based on 132 dysphonia features from sustained vowels [14]. According to the reported works on voice-based PD detection, the best results were

* Corresponding author.
E-mail addresses: gabriel.solanale@udlap.mx (G. Solana-Lavalle), roberto.rosas@udlap.mx (R. Rosas-Romero).

https://doi.org/10.1016/j.bspc.2021.102415
Received 9 October 2020; Received in revised form 26 November 2020; Accepted 10 January 2021
Available online 25 January 2021

obtained by using *support vector machine* (SVM), *random forest* (RF), *K-nearest neighbors* (KNN), or hybrid algorithms such as the one proposed by Peker et al. [8]. The most recent and relevant works on voice-based PD detection have reported a performance accuracy of 85% [6], 86% [13], 92.38% [7], 96.4% [16,17], 97.7% [18], and 98% [8,9,19,20]. It has been observed that the reported accuracy decreases as the number of subjects, within the dataset, is increased.

We have previously addressed the problem of voice-based PD detection by using the largest public dataset (64 healthy individuals and 188 PD patients), selecting a small set of features, and increasing detection accuracy [21]. However, there has not been any previous effort oriented to the provision of contextual information along with the voice-based analysis for PD detection. The binary output, from a classifying algorithm, has more clinical utility if more framing information is provided. Clinicians need to have some understanding of the basis of the recommendation and this is the most important contribution provided by the current work.

Another contribution of this work is that it is shown that gender has an influence on which features should be selected to achieve optimal discrimination between PD participants and controls [22]. PD detection, developed and applied for each gender group, has not been possible because of the reduced number of subjects within the reported datasets. We are using a dataset, where the partitioning, according to gender, is possible. Separate sets of observations, from male and female subjects, allow to find the most relevant features for each gender, without a reduction of the detection performance as Tsanas et al. explain [18]. One aim, within this study, is to obtain a better understanding of the relation among features, number of subjects, and gender.

Some challenges faced in voice-based PD detection are (1) that most works use unbalanced classes (PD patients vs. controls) and (2) a small number of observations from both classes, PD patients and controls, with the possible consequence of biased results. Almost all the most relevant and recent works, on voice-based PD detection, have reported the use of datasets with an unbalanced number of PD participants and controls: 22 PD patients vs. 30 control participants [7], 23 vs. 8 [8,9,19], 33 vs. 10 [18], 42 vs. 8 [16,17]. There are two works on voice-based PD detection that use a balanced dataset of PD participants and control participants: 20 PD patients vs. 20 controls [6] and 45 vs. 45 [20]. Since we are using the largest dataset available so far, it is possible to analyze and conduct feature selection and classification with balanced and unbalanced datasets.

The rest of the manuscript is organized as follows. The methodology section provides a description of (1) the dataset, (2) the different techniques for vocal feature extraction, (3) feature selection, (4) the classifiers used in this work, and (5) method assessment. The different feature subsets, selected by the wrappers and classification algorithms, are reported in Section 3, in Tables 5, 9 and 1. Finally, discussion of the obtained results, and conclusion are presented.

2. Methodology

The discussed approach for voice-based PD detection consists of four stages, (1) *voice recording and feature extraction* (dataset generation), (2) *feature subset selection*, (3) *classification*, and (4) *performance assessment*, which are depicted in Fig. 1. The *dataset* (first stage), used in this work, was generated at the Istanbul University and it was introduced by Sakar et al. [13]. The number of observations, within the dataset, is 756; and there are 754 features extracted from each observation by applying different signal processing techniques. To reduce the size of feature vectors and avoid the use of redundant, noisy and irrelevant attributes, *Wrappers feature subset selection* is applied. The selected and optimal features were fed to four different classifiers, *support vector machine* (SVM), *multi-layer perceptron* (MLP), *K-nearest neighbors* (KNN), and *random forest* (RF). Detection performance was measured with metrics such as accuracy and sensitivity. Finally, contextual information is also generated along with the classification of voice observations to have a better understanding of the clinical recommendation.

2.1. Dataset description

The dataset, used in this work, was introduced by Sakar et al. [13] in 2019. The group of people under study consisted of 188 PD patients (107 men and 81 women) within 65.1 years old $\pm$ 10.9 years, and 64 control individuals (23 men and 41 women) within 61.1 years old $\pm$ 8.9 years. This dataset was generated at the *Cerrahpasa Faculty of Medicine, Istanbul University*.

There were 252 individuals involved in the generation of this dataset. Each individual pronounced vowel /a/ three times (this accounts for 756 observations within the dataset), and each trial was recorded with a microphone at a sampling rate of 44.1 kHz. It has been demonstrated that the pronunciation of this vowel carry information related to the disease development such as significant changes in articulation and decrease in vowel space [23]. Furthermore, people, from different ethnic groups and native in diverse languages, are able to pronounce this vowel. The recording time is 220 seconds (9,702,000 samples per recording). The signal was divided into frames of 25 milliseconds so that these frames could be processed as stationary signals. Different signal processing techniques were used to extract 754 features from each frame. The global feature vector, representing one signal, is obtained by averaging feature vectors from all the signal frames. Signal processing and feature extraction were repeated on the 756 recordings. Different groups of features were extracted by applying six voice processing techniques: 21 *baseline features*, 11 *time frequency features*, 84 *mel frequency cepstral coefficients* (MFCC), 182 *wavelet transform based features*, 22 *vocal fold features* and 434 *tunable Q-factor wavelet transform* based features (TQWT). The Praat acoustic analysis software [24] was used by Sakar et al. to perform the steps just described [13]. This dataset is called "*Parkinson speech dataset with multiple types of sound recordings*" and it is found in the *Machine Learning Repository* of *University of California Irvine*

Table 1
Performance of the proposed method on a dataset with 384 observations, from 64 PD patients and 64 controls. Male and female subjects are included in both classes.

Feature subset selection with SVM					Feature subset selection with RF				
Classifier	Accuracy	Sensitivity	Specificity	Precision	Classifier	Accuracy	Sensitivity	Specificity	Precision
MLP	0.8698	0.8490	0.8906	0.8859	MLP	0.8750	0.9010	0.8490	0.8564
RF	0.8542	0.8281	0.8802	0.8736	RF	**0.9323**	**0.9271**	**0.9375**	**0.9368**
KNN	0.8854	0.8594	**0.9115**	0.9066	KNN	0.9167	0.9063	0.9271	0.9255
SVM	**0.8880**	**0.8646**	**0.9115**	**0.9071**	SVM	0.8229	0.8854	0.7604	0.7870
Feature subset selection with MLP					Feature subset selection with KNN				
Classifier	Accuracy	Sensitivity	Specificity	Precision	Classifier	Accuracy	Sensitivity	Specificity	Precision
MLP	**0.8646**	**0.8958**	0.8333	0.8431	MLP	0.8490	0.8646	0.8333	0.8384
RF	0.8620	0.8750	**0.8490**	**0.8528**	RF	0.9089	0.9063	0.9115	0.9110
KNN	0.7995	0.8281	0.7708	0.7833	KNN	**0.9505**	**0.9583**	**0.9427**	**0.9436**
SVM-RBF	0.8255	0.8125	0.8385	0.8342	SVM-RBF	0.8385	0.8438	0.8333	0.8351

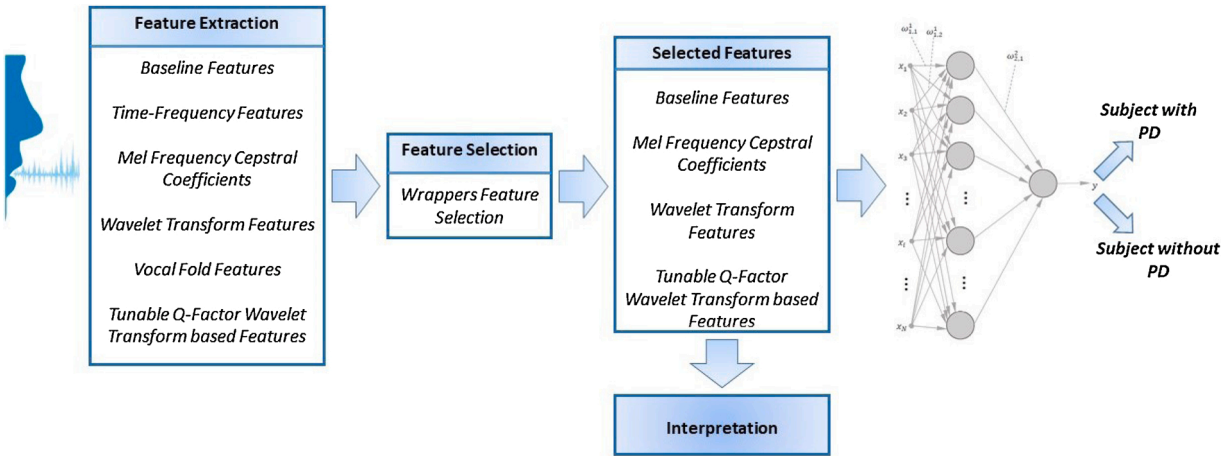

Fig. 1. Stages for voice-based PD detection.

[25].

2.2. Groups of selected features

In this work, the selected features arise from four different groups: *baseline features*, *mel frequency cepstral coefficients* (MFCC), *wavelet transform* based features (WT), and *tunable Q-factor wavelet transform* based features (TQWT). There were not features selected from the other two remaining groups: *time-frequency* features and *vocal fold* features. Features were selected by using the wrappers feature subset selection. In the following, a description of the selected groups of features is presented.

2.2.1. Baseline features

Speech features, also known as *baseline features*, have been the most important features for voice analysis and the most applied to PD detection studies; however, these features do not play a leading role in our work. Features, within this group, are known as baseline features since they have been employed as a reference when comparing the performance of other feature extraction techniques [13]. Jitter, shimmer, fundamental frequency parameters, harmonicity parameters, recurrence period density entropy (RPDE), detrended fluctuation analysis (DFA) and pitch period entropy (PPE) are the most popular baseline features [8,26,27].

2.2.2. Mel frequency cepstral coefficients

The computation of the *mel frequency cepstral coefficients* (MFCC) is based on the way the auditory system perceives differences in frequency content from voice [28,29]. The signal, under analysis, is processed with a filter bank, known as *mel bank*. The first filter is very narrow and it is used to compute how much energy exists over the lowest frequencies. At higher frequencies, the mel filters get wider. Separation between two adjacent mel filters is determined by the *mel scale* $f_{\text{mel}} = 2595\log_{10}(1+\frac{f}{700})$, which is based on the fact that the auditory system cannot discriminate closely placed frequencies and this effect gets more pronounced at higher frequencies. Furthermore, the logarithm of the energy is employed since the auditory system does not perceive voice intensity in a linear scale. Voice energy has to be amplified eight times so that the auditory perception of voice volume is doubled.

The *mel frequency cepstral coefficients* are computed by applying the *discrete cosine transform* to the *log filter bank energies*,

$$c_{\text{mel}}(n) = \sum_{m=1}^{M-1} Y(m)\,\text{Cos}\left(\frac{\pi}{M}(m-\frac{1}{2})\,n\right); n = 1, 2, \ldots, M \tag{1}$$

where $Y(m)$ is the logarithm of the weighted sum of the DFT coefficients of $x(n)$. The dynamics of the MFCC is described by the *delta* and *double-delta* coefficients. The delta coefficients are obtained by computing the difference between two MFCCs at two different frames $d_i = \frac{\sum_{j=1}^{2} j\,(c_{i+j}-c_{i-j})}{j^2}$; where d_i is the delta coefficient from frame j, and c_j is the MFCC at frame j. The double-delta coefficients are computed by replacing the mel coefficients with the delta-coefficients.

2.2.3. Discrete wavelet transform (DWT) based features

A voice signal $x(n)$ can be reconstructed by using a set of *wavelet signals*. A *wavelet* $\psi(t)$ is a short-time signal used as a building block for signal reconstruction $x(t) = \sum_j \sum_k d_{j,k}\,\psi(2^j t - k)$, where each wavelet is *scaled* $d_{j,k}\,\psi(t)$, *time shifted* $\psi(t-k)$ and *dilated* $\psi(2^j t)$. The wavelet expansion provides a *time-frequency* localization of the signal energy. The coefficient $d_{j,k}$, corresponding to component $\psi(2^j t - k)$, carries the signal energy at time position k and at time scale j. Another component, for reconstruction, is the *scaling signal* $\phi(t)$. The wavelet system, in this work, is generated from the *Haar* scaling and wavelet signals. A signal is reconstructed according to

$$x(n) = \sum_{k=-\infty}^{\infty} c_k\,\phi(t-k) + \sum_{j=0}^{\infty}\sum_{k=-\infty}^{\infty} d_{j,k}\,\psi(2^j t - k). \tag{2}$$

The set of expansion coefficients $\{c_k,\ d_{j,k}\}$ are known as the *discrete wavelet transform* (DWT). Since voice signals are characterized by short-duration high-frequency components and long-duration low-frequency components, the application of the DWT is appropriate for voice feature extraction [30]. The expansion coefficients are computed by recursive filtering and down-sampling (filter bank) [31]. The first stage of the filter bank divides the spectrum of the signal into a low-pass band and a high-pass band. The second stage divides the lower half into two quarters and so on. Each filter is characterized by the ratio $Q = \frac{\text{filter bandwidth}}{\text{filter central frequency}}$ that is constant, and this is the reason why these filters are known as "*constant-Q*".

The wavelet expansion coefficients were extracted by applying 10 levels of the DWT to the voice samples [32]. After decomposition, the energy, Shannon's and the log energy entropy, and the Teager-Kaiser energy were computed based on the DWT.

2.2.4. Tunable-Q factor wavelet transform (TQWT) features

The *tunable Q-factor wavelet transform* (TQWT) is an over-complete Wavelet Transform [27], where the *Q*-factor within the bank of band-pass filters is tuned according to the oscillatory behavior of the signal under analysis. For the case of voice signals, which are characterized by an oscillatory behavior, a relatively high *Q*-factor is appropriate. The TQWT is implemented by iteratively filtering the lower part

of a two-band signal with a *low-pass scaling filter* and a *high-pass scaling filter*, which is similar to the implementation of the DWT filtering process, but at different sampling rates [27].

Low-pass scaling modifies the incoming signal $x(n)$ by scaling the sampling frequency f_s with factor $0 < \alpha < 1$ so that the spectrum of the signal is dilated while the low-frequency content is preserved, according to

$$Y(\omega) = X(\alpha\,\omega); |\omega| < \pi. \tag{3}$$

On the other hand, *high-pass scaling* dilates the signal spectrum by a scaling factor $0 < \beta < 1$, while the high-frequency content is preserved,

$$Y(\omega) = \begin{cases} X(\beta\omega + (1-\beta)\pi) & \text{for } 0 < \omega < \pi \\ X(\beta\omega - (1-\beta)\pi) & \text{for } -\pi < \omega < 0 \end{cases} \tag{4}$$

By changing α and β, the frequency decomposition is adjusted and two parameters are defined, the *Q-factor* $Q = \frac{2-\beta}{\beta}$ and *redundancy* $r = \frac{\beta}{1-\alpha}$. During the generation of the dataset, it was found that $Q = 2$, $r = 4$, $J = 35$ (number of levels). Besides computation of the TQWT coefficients, energy/entropy values at each level were calculated.

2.3. Feature selection

The groups of features, used in this work, were selected by using a *feature subset selection* technique. A *feature subset selection* works by training a separate classifier for each possible combination of 1, 2, …, and p features. First, all the p classifiers, which are fed with one feature, are trained. All the $\frac{p(p-1)}{2}$ classifiers, which are fed with two features, are trained, and so forth. Finally, all the classification results are compared to select the best feature subset. Choosing the best subset, among $2^p - 1$ possible subsets, is computationally expensive. Thus, heuristics are applied to speed up the selection by analyzing a reduced number of feasible subsets such as the case of *feature stepwise selection* techniques: *forward stepwise selection*, *backward stepwise selection* and *wrappers feature selection*.

Forward stepwise selection begins with an empty set of features, then the most useful feature is added to the set, one by one, until all the features are included. Of all the features, which increase a set size by one, the chosen feature is the one that corresponds to the best classifier performance. At the ith step, the *i-feature subset* is the same as the $(i-1)$-*feature subset* (previous step) augmented by one feature. The best subset among all the *i-feature subsets* ($i = 1, 2, \ldots, p$) is selected. Another strategy is *backward stepwise selection* that begins with the complete set of p features, and then the least useful feature is removed, one-at-a-time. The total number of subsets, analyzed during stepwise selection techniques, amounts to $1 + \frac{p(p+1)}{2}$, which is much smaller than $2^p - 1$.

In this work, *wrappers feature subset selection* was used. It was proposed by Kohavi et al. [33]. This method is a combination of forward and backward stepwise selection. At each step, wrapper subset selection adds a new feature while it also checks the relevance of already added features. If it finds an insignificant feature, it removes that particular feature. The *Weka library* was used to run wrappers feature subset selection [34].

2.4. Classifiers

Four classification techniques are used in this work. The motivation for choosing these classifiers is that the best results, reported on previous works on voice-based PD detection, have been obtained by using these classifiers [13,7,21].

The *K-nearest neighbor* (KNN) classifies an unknown observation $\boldsymbol{x}$ by computing the Euclidean distance between $\boldsymbol{x}$ and each known observation. The k smallest distances correspond to the k nearest neighbors to $\boldsymbol{x}$. The class assigned to $\boldsymbol{x}$ is the most occurring class among the nearest neighbors.

The *multilayer perceptron* (MLP) is based on the perceptron algorithm [35]. The processing within a perceptron unit consists of combining the incoming features $\mathbf{x} = [1\, x_1\, x_2 \cdots x_p]^T$, followed by an *activation function*, according to $y = f(\omega_0 + \sum_{i=1}^{p} \omega_i x_i)$; where the coefficients $\{\omega_i; i = 0, 1, \ldots, p\}$ are the *synaptic weights*. The activation function is the *logistic function* $f(v) = \frac{1}{1+e^{-\alpha v}}$. The MLP consists of multiple layers of neurons. Each neuron's output is fully connected to all the neurons in the next layer. Training of the MLP is achieved with the *back propagation* technique, based on *gradient descent* $\omega_j(n+1) = \omega_j(n) - \eta \frac{\partial C}{\partial \omega_j}$; where C is the cost function to be minimized and η is the *learning rate*. The learning rate η affects how fast the MLP will reach a minimum in the cost function being evaluated. In this work, the learning rate was selected as 0.05.

The *support vector machine* (SVM) is a classifier that finds a hyperplane with maximum separation among different classes. The SVM output is defined as $y = b + \sum_{i=1}^{N} \omega_i K(\mathbf{x}, \mathbf{x}_i)$; where $\boldsymbol{x}$ is the observation to be classified, N is the total number of training observations, $\boldsymbol{x}_i$ is the ith *support vector*, $\{\omega_i; i = 1, 2, \ldots, N\}$ is the set of parameters to be learned along with the *bias* b, and $K(\boldsymbol{x}, \boldsymbol{x}_i)$ is a *kernel* function. In the current work, the kernel was the *Radial Basis function*, $K(\mathbf{x}, \mathbf{x}_i) = e^{-\frac{1}{2\sigma^2}||\mathbf{x}-\mathbf{x}_i||^2}$ ($\sigma = 0.1$). The architecture of the SVM consists of (1) one output (there are two classes), and (2) 8-20 input nodes. The *sequential minimal optimization* (SMO) algorithm is used to train the SVM. The convergence of the SMO is tested by checking whether the *Karush-Kuhn-Tucker* (KKT) conditions are satisfied to within a tolerance value set to 0.1.

The *random forest* (RF) is a supervised classifier that consists of a large number of *decision trees*. The feature space is partitioned into multiple regions by means of a decision tree. Region splitting is represented as a binary tree structure, where the regions correspond to leaves. The way a tree splits the feature space is quite different from others because each decision tree is trained with randomly picked observations by using *bagging* procedures. Bagging ensures small correlation values among individual decision trees. When an unknown observation is presented to a RF, each decision tree classifies the observation, and the class voted by the majority is the one taken. In this work, the RF implementation consisted of a set of 87-120 decision trees by reaching a tree depth of 150.

2.5. Training and testing sets

Each set of observations is partitioned into k fragments or *folds*, $\{C_1, C_2, \ldots, C_k\}$. $k-1$ folds are used to train a classifier, while the remaining fold is used to assess the classifier performance. This process, of choosing one fold for testing and the rest for training, is repeated k times (*k-fold cross-validation*). For each testing fold, performance metrics are measured. The final performance metrics are obtained by averaging metrics measured at each fold. For this work, $k = 10$.

2.6. Performance metrics

Evaluation metrics being used in this work are *accuracy*, *sensitivity*, *specificity*, and *precision*. When a PD patient is correctly diagnosed, it is counted as *true positive* (TP); otherwise, it will be *false negative* (FN). If a participant without PD is classified as such, it corresponds to a *true negative* (TN), otherwise it is a *false positive* (FP). The performance metrics are defined according to

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{5}$$

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{7}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

2.7. Contextual information to understand a classification outcome

To a clinician, the binary output from a black-box algorithm has limited utility, since an understanding of the association, between Parkinson disease (PD) and voice factors, is required. This is more necessary when one analysis technique gives one result, but another gives the opposite outcome. Thus, we examined the most important voice factors, which are related to PD at early stages. We searched for them by using *principal component analysis* (PCA).

PCA assigns an new representation $\mathbf{y} = [y_1, y_2, \ldots, y_p]^T$ to a *voice feature vector* $\boldsymbol{x} = [x_1, x_2, \ldots, x_p]^T$. The relationship between the voice features and a new feature is given by $y_i = \phi_{i,1}\, x_1 + \phi_{i,2}\, x_2 + \cdots + \phi_{i,p}\, x_p$, where the coefficients $\{\phi_{i,1}, \phi_{i,2}, \ldots, \phi_{i,p}\}$ are called *loadings* and the transformed features $\{y_1, y_2, \ldots, y_p\}$ are called *scores*. Whereas the loading $\phi_{i,j}$ specifies the contribution percentage of the voice feature x_j to the score y_i, the significance of the score y_i is determined by the eigenvalue λ_i, a parameter which is computed during PCA. The contribution percentage of score y_i to PD detection is computed as $\frac{\lambda_i}{\lambda_1+\lambda_2+\cdots+\lambda_p}$. Thus, a voice feature degree of influence on PD detection is found by combining the score significance (on PD detection) and a loading value, according to $\text{voice feature significance} = \frac{\phi_{i,j}\,\lambda_i}{\lambda_1+\lambda_2+\cdots+\lambda_p}$.

Once the most important voice features are known, the feature variability, for the case of PD patients, is compared with the variability, corresponding to controls, by using boxplots. A boxplot is a graph that gives a good indication of how the feature values in one population are spread out. A boxplot displays the distribution of a voice feature based on five statistical values: the minimum feature value, the first quartile Q_1 (the half between the smallest number and the median), the median value, the third quartile Q_3 (the half between the median and the highest value on the set), and the maximum feature value.

3. Results

Three sets of experiments were conducted to obtain the features with the highest contribution to PD detection along with a comparison of the variability of each feature between PD patients and controls in (1) a population with male and female subjects (balanced), (2) male subjects (balanced and unbalanced), and (3) female subjects (balanced and unbalanced).

3.1. PD detection in a population with male and female subjects

In the first set of experiments, the features with the highest contribution to PD detection were obtained by conducting experiments on a balanced dataset (the number of observations from PD subjects was the same as that from controls). Both classes included male and female subjects. Experiments, conducted on the unbalanced dataset, have been already presented [21]. To obtain a balanced dataset, some observations, from PD patients (majority), were randomly discarded giving a total of 192 observations in each class. Table 1 shows the results of using a balanced dataset of male and female subjects. Four feature subsets were selected by running wrappers with each classifier (SVM, RF, MLP, KNN). Each section in Table 1 corresponds to one of these feature subsets, and it shows the results of testing the corresponding selected feature subset with four different classifiers. The best performance results are in boldface. To demonstrate that the detection performance on balanced datasets is better than in unbalanced datasets of the same size, 384 recordings are collected randomly in an unbalanced way, and the results are shown in Table 2.

An unbalanced dataset, with 384 recordings (128 controls and 256 PD patients), was generated. One third of the dataset corresponds to recordings for controls, while two thirds correspond to PD patients. This unbalanced dataset is of the same size as the balanced one. The performance results for this unbalanced dataset are shown in Table 2. From Tables 1 and 2, it is shown that a better performance is achieved by using a balanced dataset. The results were obtained by feeding each classifier with the two best feature subsets. The best performance results are in boldface.

The first section in Table 3 shows the number of selected features from each group (baseline, MFCC, DWT, TQWT) for the case of a balanced dataset with male and female subjects. Each row shows the number of selected features from a particular group. Each column corresponds to the running of wrappers with one particular classifier (KNN, MLP, SVM, RF).

PCA is applied to all the selected features to find the four most important features for PD detection. According to the first section in Table 3, 14 features were used after feature subset selection with a KNN classifier, 9 features after selection with MLP, 14 features after selection with SVM, and 19 features after selection with RF. There were a total of 38 different features obtained by wrappers with all the classifiers. After collecting these 38 features, the PCA transformation is applied to find the four most significant *scores* $\{y_1, y_2, y_3, y_4\}$, which correspond to the four highest *eigenvalues* $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$. The *variance percentage* of each score y_i is given by $\text{variance}_i = \frac{\lambda_i}{\lambda_1+\lambda_2+\lambda_3+\lambda_4}$. The contribution percentage of each feature x_j to score y_i is given by the loading $\phi_{i,j}$. The criteria to determine the contribution of each feature x_j to PD detection is given by $\text{contribution}_j = \frac{\phi_{i,j}\,\lambda_i}{\lambda_1+\lambda_2+\lambda_3+\lambda_4}$. The four selected features, with the highest contribution to PD detection, are: (1) the *LT TKEO mean value for the eighth TQWT coefficient* with 4.93% of contribution to PD detection, (2) the *LT TKEO standard value for the sixth TQWT coefficient* with 4.873% of contribution, (3) the *LT TKEO mean value for the seventh TQWT coefficient* with 4.866% of contribution, and (4) the *det TKEO mean value for the 33rd TQWT coefficient* with 4.699% of contribution. The box plots in Fig. 2 are used to compare the variability of each feature between patients with PD and controls. The blue box shows the feature value from all the observations in PD patients, while the orange box represents the feature value corresponding to control subjects. The line in the middle of a box corresponds to the median of all feature values, known as quartile 2. The top of the box represents the median of the range of feature values between the median and the maximum (known as quartile 3). The bottom of the box is the median of the range of feature values between the minimum and the median (known as quartile 1). The cross, within a box, is the average value.

Table 4 shows two ranges of values for each feature with the highest contribution to PD detection in a mixed population. One range consists

Table 2
Performance of the proposed method on an unbalanced dataset with 384 observations, from 256 PD patients and 128 controls. Male and female subjects are included in both classes. The results were obtained by using the two best feature subsets.

Feature subset selection with KNN					Feature subset selection with RF				
Classifier	Accuracy	Sensitivity	Specificity	Precision	Classifier	Accuracy	Sensitivity	Specificity	Precision
MLP	0.8646	0.9141	0.7656	0.8864	MLP	0.8959	0.9414	0.8047	0.9060
RF	0.8984	**0.9570**	0.7813	0.8974	RF	0.9036	**0.9727**	0.7656	0.8925
KNN	**0.9167**	0.9492	**0.8516**	**0.9275**	KNN	**0.9141**	0.9492	**0.8438**	**0.9240**
SVM	0.8698	0.9141	0.7813	0.8931	SVM	0.8568	0.9336	0.7031	0.8628

Table 3
Number of features, from each group (baseline, MFCC, DWT, TQWT), within selected features. Subsets were selected by running the wrappers algorithm with four classifiers (KNN, MLP, SVM, RF).

	Balanced												Unbalanced							
	Men-Women				Women				Men				Women				Men			
	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF
Baseline	1	1	2	0	0	2	0	0	1	0	0	1	1	3	1	2	1	0	1	1
MFCC	3	3	2	2	0	1	1	3	1	0	1	1	0	2	6	2	2	3	2	3
DWT	2	0	0	5	0	2	0	1	1	1	2	2	2	0	3	2	2	1	1	2
TQWT	8	5	10	12	6	5	6	7	7	9	5	7	9	6	10	11	6	8	8	10
Total	14	9	14	**19**	6	10	7	**11**	10	10	8	**11**	12	11	**20**	17	11	12	12	**16**

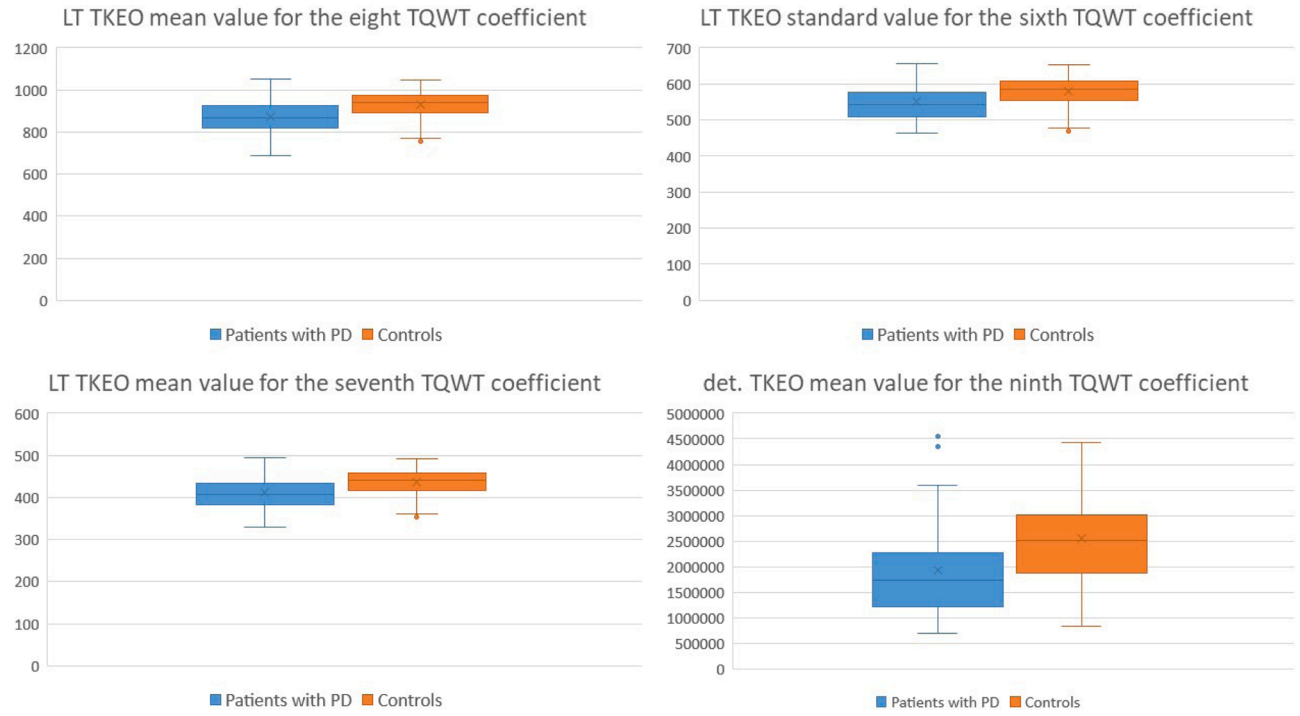

Fig. 2. Boxplots for the four features with the highest contribution to PD detection (PD patients vs. controls).

Table 4
Range of feature values for PD patients and controls in a population of men and women.

	Mixed population	
	Patient with PD	Control subject
LT TKEO mean value for the eight TQWT coefficient	$\mu = 880.8585$, $\sigma = 84.8514$ $Q_1 = 828.47$, $Q_2 = 885.33$, $Q_3 = 945.75$	$\mu = 924.423$, $\sigma = 106.8708$ $Q_1 = 882.73$, $Q_2 = 961.97$, $Q_3 = 1014.82$
LT TKEO standard value for the sixth TQWT coefficient	$\mu = 553.8791$, $\sigma = 53.18925$ $Q_1 = 516.98$, $Q_2 = 554.52$, $Q_3 = 590.9053$	$\mu = 589.0155$, $\sigma = 67.11649$ $Q_1 = 550.007$, $Q_2 = 600.9408$, $Q_3 = 633.26$
LT TKEO mean value for the seventh TQWT coefficient	$\mu = 416.53$, $\sigma = 40.001$ $Q_1 = 389.33$, $Q_2 = 414.96$, $Q_3 = 445.35$	$\mu = 443.31$, $\sigma = 50.3$ $Q_1 = 413.8$, $Q_2 = 450.88$, $Q_3 = 476.69$
det. TKEO mean value for the ninth TQWT coefficient	$\mu = 2074499$, $\sigma = 1066726$ $Q_1 = 133 \times 10^4$, $Q_2 = 187 \times 10^4$, $Q_3 = 256 \times 10^4$	$\mu = 2975794$, $\sigma = 1801452$ $Q_1 = 181 \times 10^4$, $Q_2 = 2785 \times 10^3$, $Q_3 = 37475 \times 10^2$

of average value and standard deviation. The other range consists of Quartile 1, Quartile3, and median. Each range was determined for a participants with PD and a controls.

3.2. PD detection in male subjects

In a second set of experiments, analysis, of feature contribution and variability for PD detection, was conducted on a population with only male subjects. Depending on the proportions of PD patients and controls, two cases were considered: (1) an unbalanced set of observations, with results presented in Table 5; and (2) a balanced set of observations, in Table 6. The unbalanced dataset included 321 observations from 107 male patients with PD, and 69 observations from 23 male controls. The balanced dataset included 69 observations from 23 male patients with PD, and 69 observations from 23 male controls. Four features subsets were selected by running wrappers with each classifier, and each selected subset was tested with four classifiers, which accounts for sixteen different tests for balanced and unbalanced datasets. The best results are highlighted in boldface.

An unbalanced dataset, with the same size as the balanced dataset, is generated by using 92 recording from PD male patients and 46 recordings from male controls. The classification results are shown in Table 2. These results were obtained by using the two feature selection subsets with the highest performance on the balanced dataset. It is

Table 5
Performance of PD detection in an unbalanced set of male patients with PD and controls. Four features subsets were selected by running wrappers with four different classifiers. Each selected subset was tested with four classifiers.

Feature subset selection algorithm with SVM					Feature subset selection algorithm with RF				
Classifier	Accuracy	Sensitivity	Specificity	Precision	Classifier	Accuracy	Sensitivity	Specificity	Precision
MLP	0.8692	0.9533	**0.4783**	**0.8947**	MLP	**0.9051**	0.9595	0.6522	0.9277
RF	**0.8795**	0.9813	0.4058	0.8848	RF	0.9462	0.9969	0.7101	0.9412
KNN	0.8538	0.9626	0.3478	0.8729	KNN	0.9333	0.9688	**0.7681**	**0.9511**
SVM	0.8692	**1.0000**	0.2609	0.8629	SVM	0.8333	**1.0000**	0.0580	0.8316
Feature subset selection algorithm with MLP					Feature subset selection algorithm with KNN				
Classifier	Accuracy	Sensitivity	Specificity	Precision	Classifier	Accuracy	Sensitivity	Specificity	Precision
MLP	0.8590	0.9595	**0.3913**	0.8800	MLP	0.8744	0.9346	0.5942	0.9146
RF	**0.8795**	0.9844	**0.3913**	**0.8827**	RF	0.9103	0.9782	0.5942	0.9181
KNN	0.8282	0.9315	0.3478	0.8692	KNN	**0.9436**	0.9657	**0.8406**	**0.9657**
SVM	0.8462	**0.9969**	0.1449	0.8443	SVM	0.8462	**1.0000**	0.1304	0.8425

Table 6
Performance of PD detection in a balanced set of male patients with PD and controls. Four feature subsets were selected by running wrappers with four different classifiers. Each selected subset was tested with four classifiers.

Feature subset selection algorithm with SVM					Feature subset selection algorithm with RF				
Classifier	Accuracy	Sensitivity	Specificity	Precision	Classifier	Accuracy	Sensitivity	Specificity	Precision
MLP	0.8043	0.8261	0.7826	0.7917	MLP	0.8551	0.8406	0.8696	0.8657
RF	0.8333	0.8116	0.8551	0.8485	RF	**0.9275**	**0.8841**	**0.9710**	**0.9683**
KNN	0.7971	0.8116	0.7826	0.7887	KNN	0.8478	0.7536	0.9420	0.9286
SVM	**0.8696**	**0.8406**	**0.8986**	**0.8923**	SVM	0.8188	0.8261	0.8116	0.8143
Feature subset selection algorithm with MLP					Feature subset selection algorithm with KNN				
Classifier	Accuracy	Sensitivity	Specificity	Precision	Classifier	Accuracy	Sensitivity	Specificity	Precision
MLP	**0.9058**	**0.9130**	**0.8986**	**0.9000**	MLP	0.7609	0.7826	0.7391	0.7500
RF	0.8188	0.7826	0.8551	0.8438	RF	0.8478	0.8261	0.8696	0.8636
KNN	0.8478	0.8551	0.8406	0.8429	KNN	**0.9203**	**0.8696**	**0.9710**	**0.9677**
SVM	0.7754	0.7101	0.8406	0.8167	SVM	0.7536	0.6522	0.8551	0.8182

observed that the use of a balanced set provides a better performance.

By comparing Table 5 with Table 6 and Table 6 with Table 7, it is observed that the best detection results were obtained with a balanced dataset. The third and fifth sections in Table 3 show the number of features, from different groups, selected by wrappers with four different classifiers, for the case of male balanced and unbalanced datasets. There were 34 different features used to test PD detection in men. Fig. 3 shows the boxplots for the four voice features with the highest contribution to PD detection in a male population. In Fig. 3, the variability of these features is compared between patients with PD and controls. These features, in descending order, are: (1) the *standard value for the fifth TQWT coefficient* with 5.65% of contribution to PD detection (top left panel), (2) the *Shannon entropy for the fifth TQWT coefficient* with 5.52% of contribution (top right panel), (3) the *minimum value for the fifth TQWT coefficient* with 5.43% of contribution (bottom left panel), and (4) the *TKEO mean value for the fifth TQWT coefficient* with 5.27% of contribution (bottom right panel).

Table 8 shows two ranges of values for each feature with the highest contribution to PD detection in men: (1) average value and standard deviation; (2) Quartile 1, Quartile3, and median. Each range of feature values was determined for PD patients and controls.

3.3. PD detection in female subjects

Table 9 presents the results of conducting experiments with an unbalanced dataset with 243 observations from 81 female patients with PD, and 123 observations from 41 female controls. Four feature subsets were selected by using wrappers with four classifiers. Each selected subset was tested with four classifiers. Table 10 presents the results obtained with a balanced dataset, where there are 123 observations from 41 female patients, and 123 observations from female controls.

An unbalanced dataset, which fits the size of the balanced set, is obtained by using 246 recordings from 82 female controls and 164 PD female patients. The classification results were obtained by using the two feature subsets with the best performance on the balanced dataset. It is confirmed that the use of a balanced dataset provides better results as it is shown in Table 11.

By comparing Table 9 with Table 10 and Table 10 with Table 11, the best results were obtained by using a dataset with balanced observations. The second and fourth sections in Table 3 show the number of features, from different groups, selected by wrappers for the case of balanced and unbalanced datasets of female patients and controls. There were 29 different selected features to conduct experiments of PD

Table 7
Performance results on an unbalanced dataset with 384 observations, from 256 PD patients and 128 controls. Male and female subjects are included in both classes. These results were obtained by using the two best feature subsets.

Feature subset selection with KNN					Feature subset selection with RF				
Classifier	Accuracy	Sensitivity	Specificity	Precision	Classifier	Accuracy	Sensitivity	Specificity	Precision
MLP	0.7664	0.8478	0.5435	0.7879	MLP	0.7826	0.8261	0.6957	0.8444
RF	0.8261	0.8370	0.8043	0.8953	RF	**0.8623**	**0.8913**	0.8043	0.9011
KNN	**0.8841**	**0.8804**	**0.8913**	**0.9419**	KNN	0.8478	0.8478	**0.8478**	**0.9176**
SVM	0.7463	0.9022	0.4348	0.7615	SVM	0.7536	0.8478	0.5652	0.7959

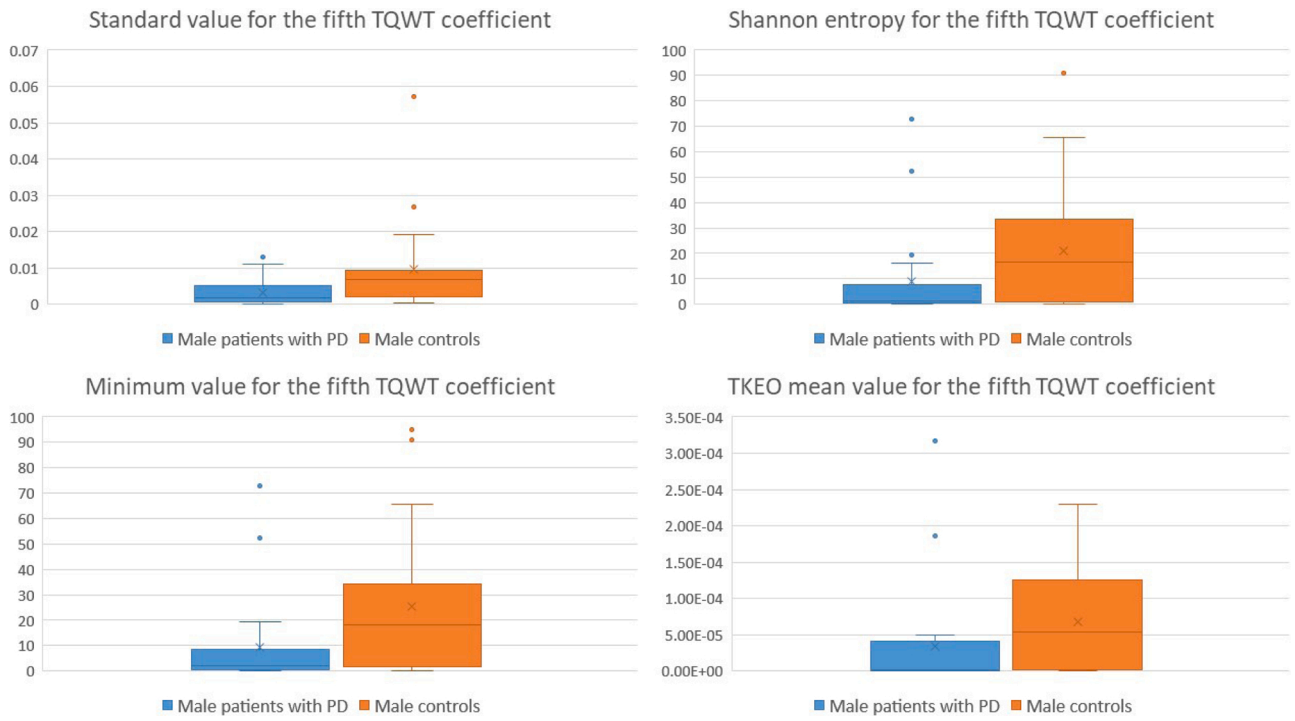

Fig. 3. Boxplots for the four features with the highest contribution to voice-based PD detection in a male population.

Table 8
Range of feature values for PD patients and controls in men.

	Male population	
	Patient with PD	Control subject
Standard value for the fifth TQWT coefficient	$\mu = 0.0027651$, $\sigma = 0.0030759$ $Q_1 = 0.0006299$, $Q_2 = 0.0014699$, $Q_3 = 0.0036567$	$\mu = 0.008074$, $\sigma = 0.010691$ $Q_1 = 0.002099$, $Q_2 = 0.005479$, $Q_3 = 0.008775$
Shannon entropy for the fifth TQWT coefficient	$\mu = 8.201105$, $\sigma = 15.56329$ $Q_1 = 0.32416$, $Q_2 = 1.2458$, $Q_3 = 6.8209$	$\mu = 17.79456$, $\sigma = 21.1533$ $Q_1 = 1.8698$, $Q_2 = 11.2268$, $Q_3 = 26.5685$
Minimum value for the fifth TQWT coefficient	$\mu = 8.07123$, $\sigma = 15.33394$ $Q_1 = 0.32416$, $Q_2 = 1.5932$, $Q_3 = 6.8209$	$\mu = 20.27424$, $\sigma = 24.97224$ $Q_1 = 1.8698$, $Q_2 = 11.3183$, $Q_3 = 28.2205$
TKEO mean value for the fifth TQWT coefficient	$\mu = 2.79 \times 10^{-5}$, $\sigma = 5.825 \times 10^{-5}$ $Q_1 = 7.115 \times 10^{-7}$, $Q_2 = 3.055 \times 10^{-6}$, $Q_3 = 2.328 \times 10^{-5}$	$\mu = 6.13 \times 10^{-5}$, $\sigma = 8.09 \times 10^{-5}$ $Q_1 = 4.22 \times 10^{-6}$, $Q_2 = 2.81 \times 10^{-5}$, $Q_3 = 8.72 \times 10^{-5}$

detection in women. After application of PCA to selected features for the female population, it was found that the four features with the highest contribution to PD detection in women were (1) the *energy of the 33rd TQWT coefficient* with 5.99% of contribution to PD detection (top left panel), (2) the *TKEO mean value for the 32nd TQWT coefficient* with 5.98% of contribution (top right panel), (3) the *Shannon entropy for the 32nd TQWT coefficient* with 5.89% of contribution (bottom left panel), and (4) the *Shannon entropy for the 33rd TQWT coefficient* with 5.83% of contribution (bottom right panel). A comparison of the variability of each feature between women with PD and controls is shown in Fig. 4.

Table 12 shows two ranges of values for each feature with the highest contribution to PD detection in women: (1) average value and standard deviation; (2) Quartile 1, Quartile3, and median. Each range of feature values was determined for a PD patient and a control subject.

4. Discussion

The use of classifiers, to assist voice-based PD detection, provides a clinician with a binary result, Positive or Negative; however, during the diagnosis, it is also important to count on appropriate contextual information for interpretation of the results. The most important factors, to interpret the detection outcome, were determined from 754 features per observation by (1) selecting the subset of the most relevant and uncorrelated features with the wrappers algorithm, followed by (2) determining the four features with the highest contribution to PD detection, among selected features, through PCA. Figs. 2–4 present a comparison of the variability of each feature between participants with PD and controls by using box plots. In almost all the twelve cases, the upper quartile, from one group (PD patient or control), and the lower quartile, from the other, do not overlap. There are ten cases where the first quartiles from both, patients and controls, do not overlap: four cases in a mixed population, four cases in a male population, and two cases in a female population. In all the cases, the range of feature values for PD patients is below the corresponding range for controls.

It is shown that the factors, associated to PD detection, are dependent on gender. According to Figs. 3 and 4, the four features with the highest contribution to PD detection are the TQWT coefficients at low-frequencies for the case of men (the fifth coefficient), and TQWT coefficients at high-frequencies for women (32nd and 33rd coefficients). According to Fig. 2, the four features, with the strongest contribution for a mixed population, are those associated to low-frequencies (sixth, seventh, eighth and ninth coefficients). These findings confirm the conclusions, obtained from studies on neuro-biology, where the analysis of structural brain imaging has shown that affected brain regions by Parkinson's disease are different between men and women [22]. A comparison of the detection performance, between male and female populations, does not show significant differences. According to Table 3, TQWT based-features are the most predominant features during detection in male and female populations with 67% and 68% of predominance, respectively.

The dataset, used in this work, was developed by Sakar et al. in 2019 [13]. In that work, feature subset selection is applied to all the groups of

Table 9
Performance of PD detection in an unbalanced set of female patients with PD and female controls. Four features subsets were selected by running wrappers with four different classifiers. Each selected subset was tested with four classifiers.

Feature subset selection algorithm with SVM					Feature subset selection algorithm with RF				
Classifier	Accuracy	Sensitivity	Specificity	Precision	Classifier	Accuracy	Sensitivity	Specificity	Precision
MLP	0.8907	0.9465	**0.7804**	**0.8949**	MLP	0.8770	0.8971	**0.8374**	**0.9160**
RF	0.8661	0.9506	0.6991	0.8619	RF	**0.9071**	**0.9630**	0.7967	0.9035
KNN	0.8606	0.9423	0.6991	0.8609	KNN	0.8552	0.8848	0.7967	0.8958
SVM	**0.8961**	**0.9711**	0.7479	0.8838	SVM	0.8388	0.9300	0.6585	0.8433
Feature subset selection algorithm with MLP					Feature subset selection algorithm with KNN				
Classifier	Accuracy	Sensitivity	Specificity	Precision	Classifier	Accuracy	Sensitivity	Specificity	Precision
MLP	**0.8880**	0.9383	**0.7886**	**0.8976**	MLP	0.8907	0.9424	0.7886	0.8980
RF	0.8634	0.9465	0.6992	0.8614	RF	0.8798	0.9424	0.7561	0.8842
KNN	0.8634	0.9095	0.7724	0.8876	KNN	**0.9590**	**0.9835**	**0.9106**	**0.9560**
SVM	0.8169	**0.9506**	0.5528	0.8077	SVM	0.8689	0.9753	0.6585	0.8495

Table 10
Performance of PD detection in a balanced set of female patients with PD and controls. Four features subsets were selected by running wrappers with four different classifiers. Each selected subset was tested with four classifiers.

Feature subset selection algorithm with SVM					Feature subset selection algorithm with RF				
Classifier	Accuracy	Sensitivity	Specificity	Precision	Classifier	Accuracy	Sensitivity	Specificity	Precision
MLP	0.7235	0.8293	0.7236	0.7500	MLP	0.7561	0.7561	0.7561	0.7561
RF	0.7520	0.7886	0.7154	0.7348	RF	**0.8862**	**0.8862**	**0.8862**	**0.8862**
KNN	0.7398	0.7317	**0.7480**	0.7438	KNN	0.8252	0.8049	0.8455	0.8390
SVM	**0.8171**	**0.8862**	**0.7480**	**0.7786**	SVM-RBF	0.7602	0.8049	0.7154	0.7388
Feature subset selection algorithm with MLP					Feature subset selection algorithm with KNN				
Classifier	Accuracy	Sensitivity	Specificity	Precision	Classifier	Accuracy	Sensitivity	Specificity	Precision
MLP	**0.8902**	**0.9106**	**0.8699**	**0.8750**	MLP	0.7642	0.8049	0.7236	0.7444
RF	0.8333	0.8374	0.8293	0.8306	RF	0.7602	0.7805	0.7398	0.7500
KNN	0.7927	0.8130	0.7724	0.7813	KNN	**0.7967**	0.8293	**0.7642**	**0.7786**
SVM	0.7846	0.7398	0.8293	0.8125	SVM	0.6911	**0.8618**	0.5203	0.6424

Table 11
Performance results on an unbalanced dataset with 246 observations, from 256 PD female patients and 128 female controls. These results were obtained by using the two best feature subsets.

Feature subset selection with KNN					Feature subset selection with RF				
Classifier	Accuracy	Sensitivity	Specificity	Precision	Classifier	Accuracy	Sensitivity	Specificity	Precision
MLP	**0.8821**	**0.9268**	0.7927	**0.8994**	MLP	0.8211	0.8598	0.7439	0.8704
RF	0.8618	0.9268	0.7317	0.8736	RF	**0.8659**	**0.9024**	**0.7927**	**0.8970**
KNN	0.8455	0.8902	0.7561	0.8795	KNN	0.8293	0.8659	0.7561	0.8765
SVM	0.8455	0.8659	**0.8049**	0.8987	SVM	0.8415	0.8841	0.7561	0.8788

features, with the result of using 50 features during PD detection. The best performance detection achieved was an accuracy of 86%. The same dataset was used later by Solana et al. in 2020 [21]. In that work, feature subset selection was applied to all the features with the result of feeding from eight to twenty selected features to the classifiers, which represents a reduction in complexity if compared with the first work. The best detection performance achieved was 94.7% of accuracy, 98.4% of sensitivity, 92.68% of specificity and 97.22% of precision. All the experiments were conducted on the 756 observations within the dataset, which includes all the male and female subjects, and an unbalanced number of participants with PD and controls. Separate experiments, for PD detection on men and women, were not conducted. An analysis to explore what factors would help physicians to understand and interpret a binary outcome is not provided. In this work, the best detection performance achieved is 95.9%of accuracy, 98.35% of sensitivity, 91.06% of specificity, and 95.6% of precision for the case of women; and 94.36% of accuracy, 100% of sensitivity, 97.1% of specificity, and 96.83% of precision for the case of men. The number of selected features fed to classifiers ranges from six to twenty.

The selected features arise from four different groups: baseline features, MFCC, DWT, and TQWT, where TQWT is the most relevant group. When features arise from only one group, the accuracy drops. Table 3 presents a case where features, from only one group (TQWT), were selected and the accuracy is dropped below 90%. On the other hand, when features come from different groups, detection accuracy considerably improves. Furthermore, the detection performance also depends on the number of observations. According to Tables 1 and 9, the best detection accuracy (95.05% and 95.9%) was obtained with the largest datasets (384 and 366 observations). According to Table 5, the detection performance, for one of the largest datasets (390 observations), reached an accuracy of 94.3%, and this may be due to the high unbalance between the 69 observations from controls, and 321 observations from PD patients.

Contrary to the results from previous studies [13,21], the KNN classifier is the algorithm, which presents the best detection performance. The KNN classifier achieves the best detection accuracy in three out of five datasets (unbalanced men, balanced men, unbalanced women, balanced women, balanced mixed). RF turned out to be the second best. RF seems to take advantage of a larger number of features in the subsets. Surprisingly, the MLP and SVM classifiers were not as good.

Fig. 4. Boxplots for the four features with the highest contribution to voice-based PD detection in a female population (PD patient vs. control participant).

Table 12
Range of feature values for PD patients and healthy individuals in women.

	Female population	
	Patient with PD	Control subject
Energy for the 33rd TQWT coefficient	$\mu = 3.73 \times 10^{-5}$, $\sigma = 3.806 \times 10^{5}$ $Q_1 = 0.00000544$, $Q_2 = 0.0000218$, $Q_3 = 0.0000566$	$\mu = 3.61 \times 10^{-5}$, $\sigma = 4 \times 10^{-5}$ $Q_1 = 7.11 \times 10^{-6}$, $Q_2 = 0.000023$, $Q_3 = 4.29 \times 10^{-5}$
TKEO mean value for the 32nd TQWT coefficient	$\mu = 3.43 \times 10^{-4}$, $\sigma = 0.0005086$ $Q_1 = 0.00003248$, $Q_2 = 0.00120245$, $Q_3 = 0.0004611$	$\mu = 5.16 \times 10^{-4}$, $\sigma = 0.000505$ $Q_1 = 0.000121$, $Q_2 = 0.00028$, $Q_3 = 0.000795$
Shannon entropy for the 32nd TQWT coefficient	$\mu = 0.4786$, $\sigma = 0.5855$ $Q_1 = 0.07624$, $Q_2 = 0.2053$, $Q_3 = 0.7244$	$\mu = 0.8426$, $\sigma = 0.7395$ $Q_1 = 0.2589$, $Q_2 = 0.5474$, $Q_3 = 1.2676$
Shannon entropy for the 33rd TQWT coefficient	$\mu = 0.4432$, $\sigma = 0.6244$ $Q_1 = 0.0465$, $Q_2 = 0.1366$, $Q_3 = 0.6183$	$\mu = 0.824$, $\sigma = 0.7489$ $Q_1 = 0.1802$, $Q_2 = 0.5902$, $Q_3 = 1.2966$

It is noticed that the detection sensitivity is higher for the case of unbalanced datasets, achieving 100% in three different subsets as it is shown in Table 5. On the contrary, detection specificity drops when unbalanced datasets are being analyzed, reaching just 5% as it is shown in Table 5. For the case of an unbalanced male dataset, good detection accuracy results were obtained, while specificity was drastically low. High specificity is preferred during PD detection, since a few false positives avoid unnecessary medication and tests for the patients.

5. Conclusion

Because of its non-intrusiveness and low economic cost, voice recordings are adequate instances to be analyzed during PD diagnosis. This work contributes to voice-based PD detection since (1) it achieves better detection performance (with low computational complexity) than previous works; (2) it is shown that features associated to PD detection are dependent on gender, a conclusion also obtained from studies on neurobiology; (3) it is shown that selected features with the highest contribution to PD detection correspond to low-frequency voice content for the case of men and high-frequency content for the case of women; (4) it is shown that a comparison, of the variability of the most important features, between participants with PD and controls, can be used as contextual information for clinical interpretation of the binary result delivered by a classifier; (5) it is confirmed that the strongest group of voice features during classification is the group of the TQWT coefficients; (6) it is also confirmed that the three most important groups of voice features, for PD detection, correspond to frequency-dependent information extracted by filter banks, a situation that is similar to the way the auditory system works; and (7) the accuracy obtained by the *k-nearest neighbor* (KNN) classifier algorithm improves considerably from previous works, reaching up to 95.9%.

Acknowledgement

The authors would like to acknowledge the support of the National Council for Research and Technology (CONACYT) in Mexico (Scholarship 934454 and stimulus 68150).

Declaration of Competing Interest

The authors report no declarations of interest.

References

[1] E. Dorsey, R. Constantinescu, J.P. Thompson, K.M. Biglan, R.G. Holloway, K. Kieburtz, F.J. Marshall, B.M. Ravina, G. Schifitto, A. Siderowf, Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030, Neurology 68 (5) (2007) 384–386.
[2] J. Michael, Fox Foundation for Parkinson Research, Parkinson's Disease Causes, 2018. Retrieved from: https://www.michaeljfox.org/understanding-parkinsons/living-with-pd.html.
[3] B. Heim, F. Krismer, R. De Marzi, K. Seppi, Magnetic resonance imaging for the diagnosis of Parkinson's disease, J. Neural Transm. 124 (8) (2017) 915–964.
[4] H. Braak, E. Ghebremedhin, U. Rüb, H. Bratzke, K. Del Tredici, Stages in the development of Parkinson's disease-related pathology, Cell Tissue Res. 318 (1) (2004) 121–134.

[5] M. Little, P. McSharry, E. Hunter, J. Spielman, L. Ramig, Suitability of dysphonia measurements for telemonitoring of Parkinson's disease, Nat. Prec. (2008).
[6] B.E. Sakar, M.E. Isenkul, C.O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, O. Kursun, Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings, IEEE J. Biomed. Health Inform. 19 (4) (2013) 828–834, https://doi.org/10.1109/JBHI.2013.2245674.
[7] D. Braga, A.M. Madureira, L. Cuelho, R. Ajith, Automatic detection of Parkinson's disease based on acoustic analysis of speech, Eng. Appl. Artif. Intell. 77 (2019) 148–158, https://doi.org/10.1016/j.engappai.2018.09.018.
[8] M. Peker, A decision support system to improve medical diagnosis using a combination of k-medoids clustering based attribute weighting and SVM, J. Med. Syst. 40 (116) (2016) 1–16, https://doi.org/10.1007/s10916-016-0477-6.
[9] H. Guruler, A novel diagnosis system for Parkinson's disease using complex-valued artificial neural network with k-means clustering feature weighting method, Neural. Comput. Appl. 28 (2017) 1657–1666, https://doi.org/10.1007/s00521-015-2142-2.
[10] L. Ali, Z. Zhang, Y. Liu, Automated detection of Parkinson's disease based on multiple types of sustained phonations using linear discriminant analysis and genetically optimized neural network, IEEE J. Transl. Eng. Health Med. 7 (2000410) (2019) 1–10, https://doi.org/10.1109/JTEHM.2019.2940900.
[11] S. Lahmiri, A. Shmuel, Detection of Parkinson's disease based on voice patterns ranking and optimized support vector machine, Biomed. Signal Process. Control 49 (2019) 427–433, https://doi.org/10.1016/j.bspc.2018.08.029.
[12] M.R. Ciucci, L.M. Grant, E.S.P. Rajamanickam, K.V. Blue, C.A. Jones, C.A. Kelm-Nelson, Early identification and treatment of communication and swallowing deficits in Parkinson disease, Semin. Speech Lang. 34 (3) (2013) 185–202, https://doi.org/10.1055/s-0033-1358367.
[13] C.O. Sakar, G. Serbes, A. Gunduz, H.C. Tunc, H. Nizam, B.E. Sakar, M. Tutuncu, T. Aydin, M.E. Isenkul, H. Apaydin, 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. 74 (2019) 255–263, https://doi.org/10.1016/j.asoc.2018.10.022.
[14] A. Tsanas, M.A. Little, P.E. McSharry, J. Spielman, L.O. Ramig, Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease, IEEE Trans. Biomed. Eng. 59 (5) (2012) 1264–1271, https://doi.org/10.1109/TBME.2012.2183367.
[15] S.A. Factor, A. Bennett, A.D. Hohler, D. Wang, J.M. Miyasaki, Quality improvement in neurology: Parkinson disease update quality measurement set, Neurology 86 (24) (2016) 2278–2283, https://doi.org/10.1212/WNL.0000000000002670.
[16] B.E. Sakar, G. Serbes, C.O. Sakar, Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease, PLOS ONE (2017) 1–18, https://doi.org/10.1371/journal.pone.0182428.
[17] S. Grover, S. Bhartia, Akshama, A. Yadav, K.R. Seeja, Predicting severity of Parkinson's disease using deep learning, Procedia Comput. Sci. 132 (2018) 1788–1794, https://doi.org/10.1016/j.procs.2018.05.154.
[18] A. Tsanas, M.A. Little, P.E. McSharry, J. Spielman, L.O. Ramig, Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease, IEEE Trans. Biomed. Eng. 59 (5) (2012) 1264–1271, https://doi.org/10.1109/TBME.2012.2183367.
[19] K. Bache, M. Lichman, UCI Machine Learning Repository, 2013. Available at http://archive.ics.uci.edu/ml.
[20] S.S. Upadhya, A.N. Sheeran, Discriminating Parkinson and healthy people using phonation and cepstral features of speech, Procedia Comput. Sci. 143 (2018) 197–202, https://doi.org/10.1016/j.procs.2018.10.376.
[21] G. Solana-Lavalle, J.C. Galan-Hernandez, R. Rosas-Romero, Automatic Parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features, Biocybern. Biomed. Eng. 40 (1) (2020) 505–516, https://doi.org/10.1016/j.bbe.2020.01.003.
[22] S.K. Yadav, N. Kathiresan, S. Mohan, G. Vasileiou, A. Singh, D. Kaura, E. R. Melhem, R.K. Gupta, E. Wang, F.M. Marincola, Gender-based analysis of cortical thickness and structural connectivity in Parkinson's disease, J. Neurol. 263 (11) (2016) 2308–2318.
[23] Y.I. Bang, K. Min, Y.H. Sohn, S.R. Cho, Acoustic characteristics of vowel sounds in patients with Parkinson disease, IEEE Trans. Biomed. Eng. 32 (3) (2013) 649–654.
[24] P. Boersma, Praat: Doing Phonetics by Computer, Ear Hear, 2011. http://www.fon.hum.uva.nl/praat/.
[25] UCI, UCI Machine Learning Repository: Parkinson Speech Dataset With Multiple Types of Sound Recordings, 2014. Available at: https://archive.ics.uci.edu/ml/machine-learning-databases/00470/.
[26] M. Peker, B. Sen, D. Delen, Computer-aided diagnosis of Parkinson's disease using complex-valued neural networks and mRMR feature selection algorithm, J. Healthc. Eng. 6 (3) (2015) 281–302, https://doi.org/10.1260/2040-2295.6.3.281.
[27] I.W. Selesnick, Wavelet transform with tunable Q-factor, IEEE Trans. Signal Process. 59 (8) (2011) 3560–3575, https://doi.org/10.1109/TSP.2011.2143711.
[28] L. Muda, M. Begam, I. Elamvazuthi, Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques, J. Comput. 2 (3) (2010) 674–693.
[29] S. Sigursson, K.B. Petersen, T. Lehn-Schiøler, Mel frequency cepstral coefficients: an evaluation of robustness of MP3 encoded music, ISMIR (2006).
[30] C.R. Torres-Santos, C.C. Cavalcante, P.C. Cortez, Wavelet transform and artificial neural networks applied to voice disorders identification, in: 2011 Third World Congress on Nature and Biologically Inspired Computing, IEEE, 2011, pp. 371–376.
[31] A. Husam, Ü. Burak-Berk, A comparative performance of discrete wavelet transform implementations using multiplierless. Wavelet Theory and Its Applications, 2018, 111.
[32] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests, IEEE Trans. Biomed. Eng. 57 (4) (2010) 884–893, https://doi.org/10.1109/TBME.2009.2036000.
[33] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997) 273–324.
[34] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, SIGKDD Explor. 11 (1) (2009) 10–18.
[35] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, Psychol. Rev. 65 (6) (1958) 386.