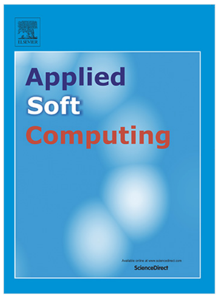
Accepted Manuscript



A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform

C. Okan Sakar, Gorkem Serbes, Aysegul Gunduz, Hunkar C. Tunc, Hatice Nizam, Betul Erdogdu Sakar, Melih Tutuncu, Tarkan Aydin, M. Erdem Isenkul, Hulya Apaydin

PII: S1568-4946(18)30579-9
DOI: https://doi.org/10.1016/j.asoc.2018.10.022
Reference: ASOC 5140

To appear in: *Applied Soft Computing Journal*

Received date : 17 April 2018
Revised date : 7 September 2018
Accepted date : 11 October 2018

Please cite this article as: 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, *Applied Soft Computing Journal* (2018), https://doi.org/10.1016/j.asoc.2018.10.022

*Highlights (for review)

- The TQWT is applied to the voice signals of Parkinson's Disease (PD) patients.
- The effectiveness of TQWT is compared with the state-of-the-art feature extraction methods.
- TQWT performed better or comparable to the state-of-the-art techniques in PD classification.
- MFCC and the TQW coefficients contain complementary information in PD classification problem.

A Comparative Analysis of Speech Signal Processing Algorithms for Parkinson's Disease Classification and The Use of The Tunable Q-Factor Wavelet Transform

C. Okan Sakar [a*], Gorkem Serbes [b*], Aysegul Gunduz [c],

Hunkar C. Tunc [a], Hatice Nizam [d], Betul Erdogdu Sakar [e], Melih Tutuncu [c],

Tarkan Aydin [a], M. Erdem Isenkul [d], Hulya Apaydin [c]

[a] Department of Computer Engineering, Bahcesehir University, Istanbul, 34353, Turkey

[b] Department of Biomedical Engineering, Yildiz Technical University, Istanbul, 34220, Turkey

[c] Department of Neurology, Cerrahpaşa Faculty of Medicine, Istanbul University, Istanbul 34098, Turkey

[d] Department of Computer Engineering, Istanbul University, Istanbul, 34320, Turkey

[e] Department of Software Engineering, Bahcesehir University, Istanbul, 34353, Turkey

e-mails: {okan.sakar@eng.bau.edu.tr; gserbes@yildiz.edu.tr;

draysegulgunduz@yahoo.com; hunkarcan.tunc@stu.bahcesehir.edu.tr; haticenizam92@gmail.com;

betul.erdogdu@eng.bau.edu.tr; tutuncumelih@yahoo.com; tarkan.aydin@eng.bau.edu.tr;

eisenkul@istanbul.edu.tr; hulyapay@istanbul.edu.tr}

Corresponding Author: C. Okan Sakar, Department of Computer Engineering, Bahcesehir University, okan.sakar@eng.bau.edu.tr

Abstract

In recent years, there has been increasing interest in the development of telediagnosis and telemonitoring systems for Parkinson's disease (PD) based on measuring the motor system disorders caused by the disease. As approximately 90% percent of PD patients exhibit some form of vocal disorders in the earlier stages of the disease, the recent PD telediagnosis studies focus on the detection of the vocal impairments from sustained vowel phonations or running speech of the subjects. In these studies, various speech signal processing algorithms have been used to extract clinically useful information for PD assessment, and the calculated features were fed to learning algorithms to construct reliable decision support systems. In this

*C. O. Sakar and G. Serbes contributed equally to this work.

study, we apply, to the best of our knowledge for the first time, the tunable Q-factor wavelet transform (TQWT) to the voice signals of PD patients for feature extraction, which has higher frequency resolution than the classical discrete wavelet transform. We compare the effectiveness of TQWT with the state-of-the-art feature extraction methods used in diagnosis of PD from vocal disorders. For this purpose, we have collected the voice recordings of 252 subjects in the context of this study and extracted multiple feature subsets from the voice recordings. The feature subsets are fed to multiple classifiers and the predictions of the classifiers are combined with ensemble learning approaches. The results show that TQWT performs better or comparable to the state-of-the-art speech signal processing techniques used in PD classification. We also find that Mel-frequency cepstral and the tunable-Q wavelet coefficients, which give the highest accuracies, contain complementary information in PD classification problem resulting in an improved system when combined using a filter feature selection technique.

Keywords: decision support system; ensemble learning; mel-frequency cepstral coefficients; parkinson's disease telemonitoring; tunable Q-factor wavelet transform.

1. Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder characterized by a large number of motor and non-motor features [1]. PD is the second – after Alzheimer – most common neurodegenerative disease seen in people over 60 [2]. The increasing prevalence rates after the age of 60 [3] and the extended life span of PD patients with the help of pharmacological or surgical interventions brings the need of accurate and reliable telemedicine systems for PD diagnosis and monitoring. Recently, many telediagnosis and telemonitoring systems have been proposed which aim to detect the disease in its early stage, decrease the number of inconvenient physical visits to the clinics for clinical examinations, and lessen the workload of clinicians [4-15].

The PD telemedicine systems are based on measuring the severity of the symptoms using non-invasive devices and tools. One of the most important symptoms seen in approximately 90% of the PD patients in the earlier stages of the disease is vocal problems. Therefore, vocal disorders-based systems constituted the focal point of the recent PD telemedicine studies [6-13]. In these studies, various speech signal processing algorithms have been used to extract clinically useful information for PD assessment, and the calculated features were fed to various learning algorithms to achieve reliable decision support systems.

The results obtained in the PD telemedicine studies showed that the choice of feature extraction and learning algorithms directly influences the accuracy and reliability of the proposed system. Many of the studies that focused on distinguishing PD patients from healthy subjects use a publicly available dataset [10] consisting of 195 sound measurements belonging to 23 PD patients and 8 healthy subjects. Another publicly available PD telediagnosis dataset [6] used in the related studies consists of multiple speech recordings of 20 PD and 20 healthy subjects. In both datasets, each speech recording is represented with similar features including vocal fundamental frequency, measures of variation in fundamental frequency, measures of variation in amplitude, measures of ratio of noise to tonal components, nonlinear dynamical complexity measures, and nonlinear measures of fundamental frequency variation. Since most of the PD telediagnosis studies perform analysis on one or both datasets, the features extracted to represent the voice signals in these datasets are the most commonly used features in the related literature. There are literature studies that proposed systems with almost 100% accuracy in discriminating healthy subjects from PD patients on these small datasets. For example, Guruler [9] combined k-means clustering based feature weighting and Complex Valued Artificial Neural Network (CVANN) for classification which reached an accuracy of 99.5%. In another study, neural networks, DMNeural, regression and decision tree algorithms were applied and an accuracy of 98.1% was proposed with neural networks classifier [13]. However, although very high classification rates have been reported on these datasets, the cross-validation techniques used in these studies cause biased results since each subject has multiple speech recordings and both training and test sets used in the experiments of these studies include the voice recordings of the same subject [6, 7]. It has already been shown that when the dataset is split properly into training/test sets using Leave-one-Subject-Out cross-validation technique, the accuracy of the proposed model dramatically decreases [7]. Besides, the number of subjects in these datasets was rather small and the accuracies obtained using complex models on such small datasets may not hold on another dataset with larger number of subjects [6, 7]. Therefore, the most commonly used features in voice-based PD telediagnosis studies, which will be referred as "baseline features" in this paper, require further analysis with larger dataset and a proper unbiased experimental setup.

Apart from the studies that used the baseline features, different feature extraction methods have been analyzed in the related domain. The most comprehensive study on the analysis of speech signal processing algorithms for the classification of PD evaluated 132 dysphonia measures grouped under 3 main feature subsets [14]. These feature subsets consist of many variations of jitter and shimmer measurements, several vocal production features built on the concept of irregular vibration of the vocal folds, quantification of noise, estimations of signal-to-noise ratio, and Mel-frequency cepstral coefficients. However, in [14] rather than analyzing the performance of each signal processing technique individually, the features extracted using various signal processing techniques are merged into a single feature set and a feature selection process has been performed on this single feature set. In our study, we use a different feature subset categorization. This approach has two main reasons. First, we aim to analyze and present the performance of each feature subset individually in distinguishing the healthy subjects from PD patients using different classifiers. Second, we prefer to analyze the baseline features as a separate group and compare the performance of each feature subset with that of this group since baseline feature set represents the most commonly used feature set in this domain.

In the proposed study, to the best of our knowledge for the first time, we also apply the tunable Q-factor wavelet transform (TQWT), which has higher frequency resolution than classical dyadic discrete wavelet transform [15, 16], to the voice signals of PD patients with the aim of extracting discriminative features. In the TQWT, the frequency selectivity of the band-pass filters used in decomposition and reconstruction stages can be determined by changing certain parameters named as Q (Q-factor of band-pass filters), r (oversampling rate or redundancy) and J (the number of analysis and synthesis levels). Thus, an optimum time-frequency representation can be obtained by choosing the optimal Q-factor, which is defined as the ratio of the center frequency of the wavelet to its bandwidth, according to the characteristic of the processed signal. It is seen that when relatively higher Q values are employed in the TQWT analysis, this paves the way for obtaining narrower frequency responses allowing to obtain much better decomposition of subbands to span the frequency range. Various kind of overcomplete wavelet transforms (like the TQWT), in which the frequency resolution of the wavelets can be adjusted, were successfully used in various audio signal processing applications such as time scaling of audio signals, tone scrolling [17], decomposition of audio signals into components [18] and restoration of audio signals [19]. These studies showed that the audio signals having oscillatory components can be represented in the time-frequency axis in an optimum manner when the proper *Q* values, which fits the nature of processed signal, are set. The main motivation of employing TQWT in our study is that by tuning the Q-factor of the wavelet functions to unveil the characteristic behavior of the healthy and PD speech samples, more efficient and robust time-scale representations can be obtained. Hence, the TQWT is proposed as the novel feature extractor in our study to catch the distinctive changes in the time-frequency axis between the normal and pathologic individuals.

Another important contribution of this study is the comparison of the signal processing methods using ensemble learning approaches which combines the predictions of seven classifiers. The ensembles consists of the most commonly used classifiers in the domain of dysphonia-based PD telediagnosis systems which are support vector machines (SVM) with linear and RBF kernels, Multilayer Perceptron, a probabilistic classifier (Naive Bayes), logistic regression, a decision-tree based learning algorithm (Random Forest) and an instance-based learning algorithm (k-Nearest Neighbor). By combining the predictions of multiple classifiers, we aim to decrease the effect of classifiers in the comparison of feature extraction methods and also the variance of the final classification models. Additionally, we rank the features according to their relevance with the class label and redundancy with the other features using the minimum Redundancy-Maximum Relevance [20] filter feature selection method and present the results obtained using the top-ranked features.

2. Materials and Methods

2.1.Dataset Description

The data used in this study were gathered from 188 patients with PD (107 men and 81 women) with ages ranging from 33 to 87 (65.1±10.9) at the Department of Neurology in Cerrahpaşa Faculty of Medicine, Istanbul University. The control group consists of 64 healthy individuals (23 men and 41 women) with ages varying between 41 and 82 (61.1±8.9). During the data collection process, the microphone is set to 44.1 KHz and following the physician's examination, the sustained phonation of the vowel /a/ was collected from each subject with three repetitions. All subjects were informed about the data collection process, signed informed consent, and attended the test voluntarily in accordance with the approval of Clinical Research Ethics Committee of Bahcesehir University.

2.2.Feature Extraction

Idiopathic PD is a neurodegenerative disorder which occurs due to loss of neuromelanin-containing neurons in substantia nigra pars compacta in the midbrain which leads to decrease of striatal dopamine. In literature [21-22], it was shown that PD affects speech even at an early stage, and therefore speech features have been successfully employed to assess PD and monitor its evolution after surgical or pharmacological treatment. 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 speech features used in PD studies [10-14]. In our study, these features are referred as "baseline features" and employed for comparing the performance of the other feature extraction methods analyzed in this study. Except from the RPDE, DFA and PPE, the Praat [23] acoustic analysis software is utilized for extracting baseline features and the detailed expressions of baseline features are given in Table 1. Speech intensity, formant frequencies and bandwidth-based features are also extracted from the spectrograms of the speech signals by using the Praat, and detailed explanations of these three feature subsets are also given in Table 1.

In literature, Mel-Frequency Cepstral Coefficients (MFCCs), which emulate the effective filtering properties of the human ear, have been used as a robust feature extraction method in the context of speaker identification, automatic speech recognition, biomedical voice assessment and Parkinson's disease diagnosis [14, 24, 25]. In MFCC extraction method, cepstral analysis is combined with spectral domain partitioning by using triangular shape overlapped filter-banks and this results in a narrow spectral sampling. In PD studies, MFCCs are employed in detecting subtle changes in the motion of the articulators (tongue, lips) which are known to be affected in PD [14]. In this study, mean and standard deviation of the original 13 MFCCS plus log-energy of the signal and their first-second derivatives are employed as features resulting in 84 features which are given in Table 1.

In [26], as a novel approach in PD studies, wavelet transform (WT) based features that were obtained from the raw fundamental frequency (F_0) contour of speech samples were employed as the indicators of Unified Parkinson's Disease Rating Scale (UPDRS). The idea behind the usage of WT based features was that the deviation from the exact periodicity of a sustained vowel would be minimum for the healthy speech samples while there would be significant deviations in pathological speech samples [27]. In our study, to quantify the performance of WT based features, which are obtained from the raw F_0 contour and also from the log transform of the F_0 contour, as suggested in [28], 10-levels discrete wavelet transform is applied to speech samples. After decomposition, the energy, Shannon's and the log energy entropy, and the Teager-Kaiser energy of both the approximation and detailed coefficients are calculated resulting in 182 WT features related with F_0 as shown in Table 1.

Additional to these features, in this study, features based on vocal fold vibration pattern and the effects of noise on vocal fold are extracted to quantify their success and to compare them with the proposed TQWT based features. In this context, the Glottis Quotient (GQ), Glottal to Noise Excitation (GNE), Vocal Fold Excitation Ratio (VFER) and Empirical Mode Decomposition (EMD) features are calculated and the explanation of these features, which are named as vocal fold features for simplicity, can be found in Table 1.

We utilized the Voice Analysis Toolbox [4, 8, 26] for extracting the RPDE, DFA, PPE, MFCCs, WT based features and vocal fold features.

An important contribution of this study is to employ the tunable Q-factor wavelet transform (TQWT), which is a fully discrete and over-complete WT, as the main feature extractor [15]. In the TQWT, the *Q*-factor of the wavelets can easily be tuned according to the behavior of signal to which it is applied. For the analysis of speech signals, having oscillatory time domain behavior, a relatively high *Q*-factor transform would be more appropriate, whilst a low *Q*-factor transform would give better results in the process of non-oscillatory signals. The TQWT consists of two channel filter-banks, which are iteratively applied, and in each iteration, the low-pass filter output is given to the next iteration low/high pass filters as inputs. At the end of decomposition stage, considering the J as the number of levels, there will be J+1 subbands coming from J high-pass filter and one final low-pass filter outputs. The third parameter r is named as the redundancy or oversampling rate which controls the undesired excessive ringing in order to localize the wavelets in time domain without affecting their shape.

Table 1. Overview of the feature sets used in this study (except TQWT)

Feature Set	Measure	Explanation	# of features
Baseline Features	Jitter variants	Jitter variants are employed to capture the instabilities occurred in the oscillating pattern of the vocal folds and this feature sub-set quantifies the cycle-to-cycle changes in the fundamental frequency.	5
	Shimmer variants	Shimmer variants are also employed to capture instabilities of the oscillating pattern of the vocal folds, but this time this feature sub-set quantifies the cycle-to-cycle changes in the amplitude	6
	Fundamental frequency parameters	The frequency of vocal fold vibration. Mean, median, standard deviation, minimum and maximum values were used.	5
	Harmonicity parameters	Due to incomplete vocal fold closure, increased noise components occur in speech pathologies. Harmonics to Noise Ratio and Noise to Harmonics Ratio parameters, which quantify the ratio of signal information over noise, were used as features.	2
	Recurrence Period Density Entropy (RPDE)	RPDE gives information about the ability of the vocal folds to sustain stable vocal fold oscillations and it quantifies the deviations from F_0.	1
	Detrended Fluctuation Analysis (DFA)	DFA quantifies the stochastic self-similarity of the turbulent noise.	1
	Pitch Period Entropy (PPE)	PPE measures the impaired control of fundamental frequency F_0 by using logarithmic scale.	1
Time Frequency Features	Intensity Parameters	Intensity is related with the power of speech signal in dB. Mean, minimum and maximum intensity values were used.	3
	Formant Frequencies	Frequencies amplified by the vocal tract, the first four formants were used as features.	4
	Bandwidth	The frequency range between the formant frequencies, the first four bandwidths were employed as features.	4
Mel Frequency Cepstral Coefficients (MFCCs)	MFCCs	MFCCs are employed to catch the PD affects in vocal tract separately from the vocal folds	84
Wavelet Transform based Features	Wavelet transform (WT) features related with F_0	WT features quantify the deviations in F_0	182
Vocal Fold Features	Glottis Quotient (GQ)	GQ gives information about opening and closing durations of the glottis. It is a measure of periodicity in glottis movements.	3
	Glottal to Noise Excitation (GNE)	GNE quantifies the extent of turbulent noise, which caused by incomplete vocal fold closure, in the speech signal.	6
	Vocal Fold Excitation Ratio (VFER)	VFER quantifies the amount of noise produced due to the pathological vocal fold vibration by using nonlinear energy and entropy concepts.	7
	Empirical Mode Decomposition (EMD)	EMD decomposes a speech signal into elementary signal components by using adaptive basis functions and energy/entropy values obtained from these components are used to quantify noise.	6

In the TQWT, filters having rational transfer functions, which are computationally efficient and can be specified in frequency domain, are employed in decomposition and reconstruction stages. The TQWT wavelets have constant Q-factor, which is defined beforehand, during the transformation process and the transform inherits the perfect reconstruction property that makes it a perfect tool for signal manipulations like denoising, compression, etc. The two channel filterbank structure of the TQWT for a single level can be seen in Figure 1. As it can be seen, for a single level transform, the input signal $s[n]$ is decomposed into low-pass subband signal $c^0[n]$ and high-pass subband signal $d^1[n]$ with sampling rate f_s resulting in sampling frequencies αf_s and βf_s respectively. $H_0(\omega)$ and $H_1(\omega)$ are the frequency responses of low-pass and high-pass filters respectively, and they can be defined mathematically for the j^{th} level as follows:

$$H_0^{(j)}(\omega) = \begin{cases} \prod_{m=0}^{j-1} H_0\left(\frac{\omega}{\alpha^m}\right), & |\omega| \le \alpha^j \pi \\ 0, & \alpha^j \pi \le |\omega| \le \pi \end{cases} \quad (1)$$

and

$$H_1^{(j)}(\omega) = \begin{cases} H_1\left(\frac{\omega}{\alpha^{j-1}}\right) \prod_{m=0}^{j-2} H_0\left(\frac{\omega}{\alpha^m}\right), & (1-\beta)\alpha^{j-1}\pi \le |\omega| \le \alpha^{j-1}\pi \\ 0, & for\ others\ \ \omega \in [-\pi, \pi] \end{cases} \quad (2)$$

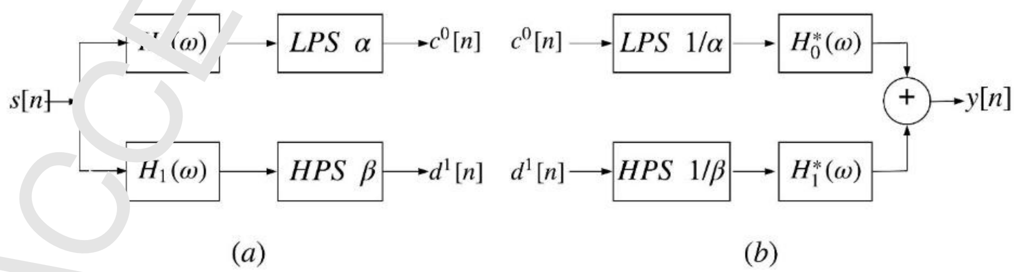


Figure 1. Decomposition (a) and reconstruction (b) stages of single level TQWT. LPS and HPS represent low-pass scaling and high-pass scaling respectively

The equivalent system for the j^{th} level TQWT decomposition for the input signal $s[n]$ and the generated low-pass/high-pass subband signals $c^j[n]/\,d^j[n]$ are given in Figure 2. In the TQWT, by using the scaling parameters α and β, the redundancy and Q parameters can be expressed as:

$$r = \frac{\beta}{1-\alpha}\,, Q = \frac{2-\beta}{\beta} \qquad (3)$$

In the proposed study, sustained phonations of the vowel /a/ are collected from healthy and PD subjects with the aim of constructing mathematical models that can be used in classification. The used healthy speech signals have oscillatory characteristics due to nearly periodic vocal fold vibration pattern whereas this nearly periodic pattern is distorted in PD speech signals. Therefore, in this study, the parameters of the TQWT are tuned in accordance with the time domain characteristics of speech signals. This tuned parameter set yields improved frequency resolution in the transform and more discriminative ability for the model. In the TQWT algorithm, the performance of the transform relies on Q (Q-factor), r (redundancy) and J (number of levels) parameters. It was stated in [15] that the value of r must be chosen as equal or greater than 3 for preventing the undesired ringings in wavelets. Therefore, after numerous trial and error experiments, the suitable values of r were chosen as 3, 4 and 5. Additionally, in order to find the suitable number of analysis levels, J values were tested between 5 and 50 for different Q values and best accuracy values were found for the suitable Q-r pairs. As mentioned before, Q value controls the oscillatory behavior of wavelets and to obtain more oscillatory wavelets higher Q values must be employed. During the experiments, reasonable Q values changing from 1 to 10 in step of one were tested. It was seen that when higher Q values are chosen, greater number of levels (J values) must also need to be employed in order to span the entire frequency axis. After testing period, best parameter set giving the highest accuracy was found as Q=2,r=4,J=35 in which a relatively higher Q-value was employed. Additionally, the parameters were also set to Q=1, r=3, J=8 and Q=4, r=5, J=45 in order to show the performance of the TQWT when the Q parameter is not chosen properly.

In some cases, the sparse representation of wavelet coefficients for the input signal can be more useful. Hence, the coefficient set obtained with the TQWT from the healthy and PD speech samples were processed with Basis Pursuit (BP) approach to obtain sparse representations [15]. In BP, to get the sparse representation, the below optimization problem must be solved:

$$\min_a \|a\|_1 \;\; \text{such that} \;\; \Phi a = x \qquad (4)$$

where Φ is a matrix which represents decomposition functions of the TQWT and a is the wavelet coefficients vector. After both the normal and sparse decompositions with the TQWT, energy/entropy values of each level were calculated, and these energy/entropy values were employed in healthy/PD subject classification as features.

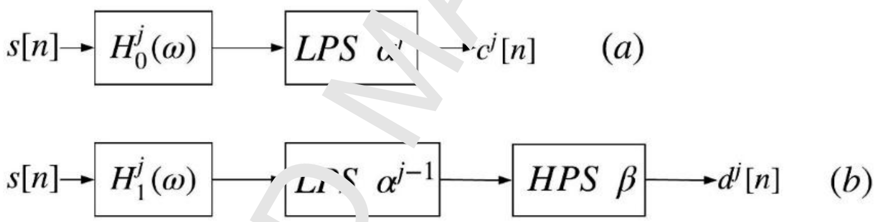


Figure 2. The equivalent system for the j^{th} level TQWT decomposition. $c^j[n]$ is the low-pass subband signal (a) and $d^j[n]$ is the high-pass subband signal (b)

2.3.Classification

Following the feature extraction step, the obtained feature vectors are standardized so that each feature has a zero mean and unit variance. Then the feature subsets are fed into multiple classifiers to discriminate healthy subjects from PD patients. We also present the results obtained with the TQWT with different Q values. The aim is to evaluate the performance of each feature extraction method in dysphonia-based PD telediagnosis systems. We use leave-one-subject-out (LOSO) cross-validation technique to validate the generalization ability of the classification models. The LOSO method is based on leaving the voice sample of one individual out to be used for validation as if it is an unseen individual and using the voice samples of the other subjects for training. Since every subject provided three recording samples, we performed a majority voting on the predictions of these samples to make the final prediction on a subject.

The standardized features are fed to SVMs with linear and RBF kernels, Multilayer Perceptron, Naive Bayes, Logistic Regression, Random Forest and k-NN algorithms. We use overall accuracy, F1-score and Matthew's correlation coefficient metrics to compare the results of various feature subsets and classifiers.

2.4.Ensemble of Classifiers

In ensemble classification, multiple individual learners are trained on the same classification task and the class predictions are combined using a combination strategy. The theoretical [30, 31], and empirical [32, 33] findings on ensemble learning showed that in order to obtain a better predictive final model, the individual members of the ensemble must be diverse and accurate. Based on this finding, in this study we use different types of learning algorithms in our ensemble model which are SVMs with linear and RBF kernels, Multilayer Perceptron, a probabilistic classifier (Naive Bayes), Logistic Regression, a decision-tree based learning algorithm (Random Forest), an instance-based learning algorithm (k-NN).

The predictions of individual classifiers are combined using voting or stacking strategies. The final prediction of an ensemble consisting of M members is:

$$y = \sum_{i=1}^{M} w_i d_i \qquad (5)$$

satisfying

$$w_i \geq 0, \text{and} \sum_{i=1}^{M} w_i = 1 \quad (6)$$

where w_i is the weight of the prediction of the i^{th} ensemble member and d_i represents the prediction of the i^{th} ensemble member. We apply simple majority voting and stacking with linear kernel SVM in our experiments.

2.5.Feature Ranking

We use the minimum redundancy-maximum relevance (mRMR) [20] based filter feature selection method to determine the most effective features and also to obtain a more robust and accurate PD classification model by reducing the high dimensional problem to a minimum set with maximum joint dependency. The mRMR approach is based on maximizing the joint dependency of top ranking variables on the target variable by reducing the redundancy among them [20, 34]. The mRMR algorithm has been successfully applied to a wide variety of machine learning problems as a preprocessing step including gene expression analysis [35], protein structure prediction [36], biomedical decision support systems [12, 13, 14], hyperspectral data classification [37] and churn prediction [38]. Therefore, we also apply mRMR to evaluate the effectiveness of feature subsets and obtain a minimal set of features to separate PD patients and healthy subjects. The overview of the proposed end-to-end PD classification system is shown in Figure 3.

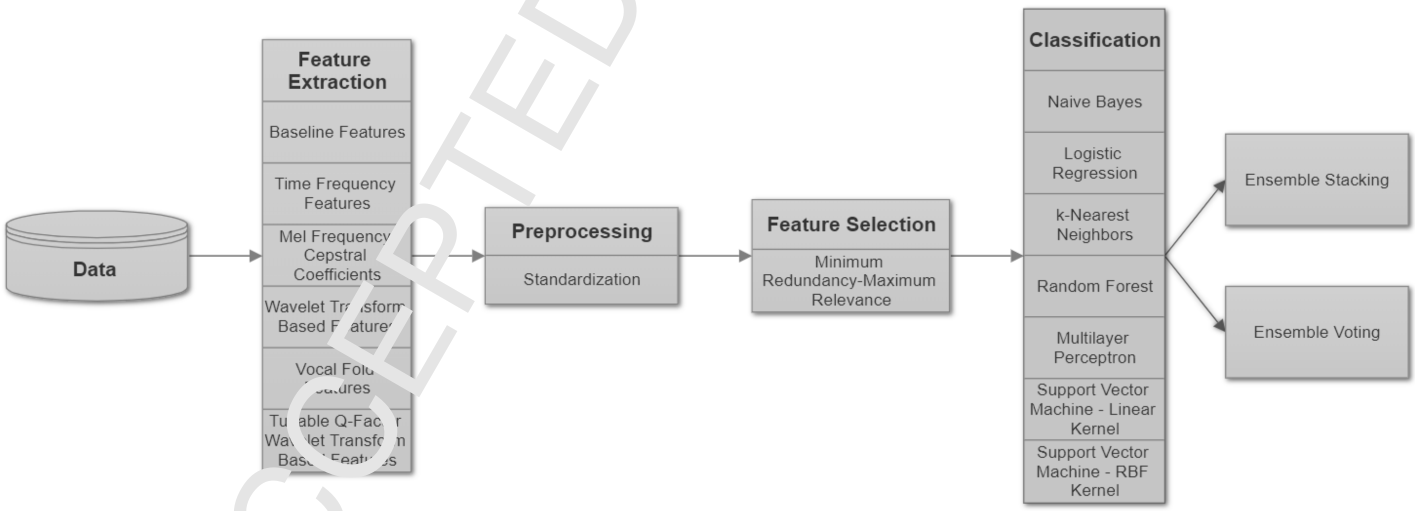


Figure 3. Overview of the proposed PD classification system

2.6.Statistical Significance of the Results

We perform McNemar's test to examine whether the differences between the prediction performances of feature subsets are statistically significant or not [29]. This test is used in dichotomous classification to identify whether two algorithms have the same error rate or not. In this test, after obtaining the predictions of two classifiers, the number of samples misclassified by both (e_{00}), by the first algorithm but not the second (e_{01}), by the second algorithm but not first (e_{10}) and the number of samples correctly classified by both (e_{11}) are calculated. Then, these values are placed to a 2x2 contingency table. The null hypothesis is that the classification algorithms have the same error rate, and, in that case, we expect $e_{01} = e_{10}$ [29]. A chi-square statistic with one degree of freedom is worked out by the formula shown below to test this null hypothesis:

$$\frac{(|e_{01}-e_{10}|-1)^2}{e_{01}+e_{10}} \sim X_1^2 \quad (7)$$

If the value of X_1^2 is less than $X_{\alpha,1}^2$, the two algorithms are considered to have the same error rate. Otherwise, the null hypothesis is rejected at significance level α. For α = 0.05, $X_{0.05,1}^2 = 3.84$.

3. Experimental Results

3.1.Classification Performance of Individual Feature Subsets

Table 2 shows the test set accuracies, F1-scores, and Matthew's correlation coefficients obtained with each individual feature subset. The average performances of the classifiers and the ensemble learning accuracies obtained by combining the predictions of classifiers using voting and stacking approaches are also presented in Table 2. The highest accuracy of 0.84 with 0.83 F1-score and 0.54 MCC is achieved by feeding MFCCs to SVM-RBF classifier. It is seen that the ensemble approach that combines the predictions of MFCC based classification models with voting and stacking strategies did not improve the SVM-RBF results. McNemar's test revealed that the highest performed SVM-RBF model which uses the MFCC features does not perform significantly better than the logistic regression model which has the highest performance based on baseline features at significance level 0.05 (X_1^2= 3.69).

Table 2. Results obtained with individual feature subsets

	Baseline Features			**MFCC**			**Wavelet Features Extracted from F_0**		
	Accuracy	**F1-Score**	**MCC**	**Accuracy**	**F1-Score**	**MCC**	**Accuracy**	**F1-Score**	**MCC**
Naive Bayes	0.53	0.55	0.21	0.56	[illegible]	0.31	0.72	0.71	0.24
Logistic Regression	**0.79**	**0.75**	**0.34**	0.83	0.82	0.52	0.76	0.72	0.25
k-NN	0.75	0.71	0.22	0.80	0.77	0.40	0.73	0.71	0.22
Multilayer Perceptron	0.77	0.75	0.32	0.82	0.81	0.49	**0.78**	**0.74**	**0.31**
Random Forest	0.77	0.75	0.31	[illegible]	0.80	0.49	0.77	0.75	0.31
SVM (Linear)	0.77	0.72	0.28	0.81	0.80	0.47	0.75	0.69	0.17
SVM (RBF)	0.77	0.74	0.29	**0.84**	**0.83**	**0.54**	0.77	0.72	0.25
Average	0.73	0.71	0.29	0.78	0.76	0.45	0.75	0.72	0.25
Std. Dev.	0.10	0.08	0.06	0.11	0.09	0.08	0.02	0.02	0.05
Ensemble with Voting	0.79	0.75	0.34	0.84	0.83	0.53	0.75	0.70	0.19
Ensemble with Stacking	0.78	0.75	[illegible]	0.83	0.82	0.52	0.77	0.74	0.29
	Bandwidth + Formant			**Intensity-Based**			**Vocal Fold-Based**		
	Accuracy	**F1-Score**	**MCC**	**Accuracy**	**F1-Score**	**MCC**	**Accuracy**	**F1-Score**	**MCC**
Naive Bayes	0.74	0.69	0.15	0.57	0.59	0.29	0.69	0.70	0.26
Logistic Regression	**0.77**	[illegible]	**0.25**	0.74	0.64	-0.04	0.76	0.72	0.25
k-NN	0.76	0.71	0.23	0.75	0.72	0.24	0.76	0.71	0.23
Multilayer Perceptron	0.76	0.73	0.25	0.76	0.67	0.22	0.75	0.72	0.23
Random Forest	0.75	0.71	0.21	**0.77**	**0.74**	**0.30**	**0.77**	**0.74**	**0.30**
SVM (Linear)	0.75	0.64	0.0	0.75	0.64	0.0	0.76	0.68	0.18
SVM (RBF)	0.77	0.71	0.25	0.75	0.64	0.0	0.77	0.72	0.25
Average	0.76	0.70	0.18	0.72	0.67	0.17	0.75	0.71	0.24
Std. Dev.	0.01	0.03	0.10	0.08	0.06	0.15	0.03	0.02	0.04
Ensemble with Voting	0.76	0.70	0.22	0.75	0.65	0.11	0.76	0.72	0.25
Ensemble with Stacking	0.73	0.68	0.13	0.76	0.74	0.29	0.77	0.74	0.30

3.2. *Proposed Classification Method with the TQWT Based Features*

The normal and sparse TQWT results with two different Q-factor values are shown in Table 3. It is seen that the highest individual classifier accuracy of 0.85 with 0.84 F1-Score and 0.57 MCC is obtained by feeding normal Q-wavelet transform features, which are extracted with the relatively high Q-factor (selected as 2) analysis, into the multilayer perceptron classifier. It should be noted that increasing the Q value too much (such as 4,5,6 and etc.) does not increase the classification accuracy due to the need of excessive number of analysis levels (J value). When the number of decomposition levels become huge, the ratio of overlapped information between subbands also dramatically increases and the increasing number of features containing redundant information causes curse of dimensionality problem in the learning step.

Table 3. Results obtained with the TQWT feature extraction method (energy and entropy-based features)

	Normal: *Q*=1, *r*=3, *J*=8			**Sparse: *Q*=1, *r*=3, *J*=8**		
	Accuracy	**F1-Score**	**MCC**	**Accuracy**	**F1-Score**	**MCC**
Naive Bayes	0.66	0.68	0.25	0.75	0.76	0.36
Logistic Regression	0.75	0.73	0.27	**0.81**	**0.80**	**0.46**
k-NN	0.78	0.74	0.32	0.80	0.77	0.39
Multilayer Perceptron	0.78	0.77	0.39	0.75	0.74	0.28
Random Forest	0.79	0.76	0.36	0.81	0.78	0.42
SVM (Linear)	0.77	0.76	0.34	0.80	0.79	0.44
SVM (RBF)	**0.81**	**0.78**	**0.42**	0.81	0.78	0.45
Average	0.76	0.74	0.34	0.79	0.77	0.39
Std. Dev.	0.05	0.03	0.04	0.03	0.02	0.07
Ensemble with Voting	0.80	0.78	0.40	0.81	0.79	0.45
Ensemble with Stacking	0.80	0.77	0.40	0.79	0.75	0.34
	Normal: *Q*=2, *r*=4, *J*=35			**Sparse: *Q*=2, *r*=4, *J*=35**		
	Accuracy	**F1-Score**	**MCC**	**Accuracy**	**F1-Score**	**MCC**
Naive Bayes	0.73	0.74	0.39	0.78	0.78	0.43
Logistic Regression	0.79	0.79	0.45	0.72	0.72	0.28
k-NN	0.83	0.81	0.50	0.81	0.79	0.45
Multilayer Perceptron	**0.85**	**0.84**	**0.57**	0.78	0.77	0.39
Random Forest	0.83	0.82	0.52	**0.82**	**0.80**	**0.47**
SVM (Linear)	0.79	0.79	0.45	0.70	0.71	0.23
SVM (RBF)	0.85	0.83	0.56	0.80	0.78	0.41
Average	0.80	0.80	0.48	0.77	0.76	0.38
Std. Dev.	0.04	0.03	0.07	0.05	0.04	0.10
Ensemble with Voting	0.85	0.84	0.57	0.79	0.78	0.4
Ensemble with Stacking	0.84	0.83	0.54	0.81	0.80	0.47
	Normal: *Q*=4, *r*=5, *J*=45			**Sparse: *Q*=4, *r*=5, *J*=45**		
	Accuracy	**F1-Score**	**MCC**	**Accuracy**	**F1-Score**	**MCC**
Naive Bayes	0.64	0.66	0.26	0.69	0.71	0.32
Logistic Regression	0.77	0.77	0.38	0.75	0.75	0.33
k-NN	0.79	0.76	0.37	**0.80**	**0.78**	**0.41**

Multilayer Perceptron	0.79	0.78	0.41	0.77	0.77	[illegible]
Random Forest	**0.81**	**0.80**	**0.45**	**0.80**	**0.78**	**0.41**
SVM (Linear)	0.75	0.75	0.34	0.71	[illegible]	0.23
SVM (RBF)	0.80	0.76	0.38	0.77	0.73	0.27
Average	0.76	0.75	0.37	0.75	0.75	0.35
Std. Dev.	0.06	0.05	0.06	0.05	[illegible]	0.07
Ensemble with Voting	**0.81**	**0.80**	**0.45**	0.76	0.75	0.31
Ensemble with Stacking	0.80	0.79	0.43	0.77	0.76	0.34

The highest individual, average and both voting and stacking ensemble accuracies obtained with the best setting of the TQWT are higher than or equal to that of the feature extraction methods given in Table 1. We performed McNemar's test to assess whether the error rate obtained with the best setting of TQWT is significantly lower than the error rate obtained with the best performed feature subset in Table 1, MFCC, and baseline features. The statistical test revealed that the error rate of the best model obtained with the TQWT features is significantly less than the highest performed model based on baseline features (X_1^2= 4.55), but not statistically different from the highest performed model based on MFCC features (X_1^2= 0.09). These results show that the TQWT features are effective in discriminating PD patients from healthy subjects and could be used in dysphonia-based PD telediagnosis systems.

3.3.Combining Feature Sets and Classification with Top-Ranked Features

We selected the top-50 features by applying the mRMR feature selection algorithm to the combination of all feature subsets. The feature selection step was conducted on the training data at each step to prevent feature subset selection bias. Table 4 demonstrates the average number of selected features from each feature subset. We present the accuracy resulted by the combination of all feature subsets in Table 5 along with the results with holding out the TQWT and MFCC subsets. As seen in Table 5, the highest metrics achieved were 0.86 accuracy, 0.84 F1-score and 0.59 MCC with the SVM-RBF classifier by using the top-50 features selected by mRMR on the combination of all feature subsets.

Table 4. Average distribution of the top-50 features selected by the mRMR filter.

	All Feature Subsets Except TQWT	All Feature Subsets Except MFCC	All Feature Subsets
Baseline (n=26)	5	5	4
Intensity (n=3)	1	1	0
Bandwidth + Formant (n=8)	5	2	2
MFCC (n=84)	27	-	10
WT applied to F_0 (n=182)	4	1	1
Vocal Fold (n=22)	8	4	3
TQWT(n=432)	-	37	30

Table 5. Results obtained with top-50 features selected by the mRMR filter on the combined feature subsets.

	All Feature Subsets Except TQWT			All Feature Subsets Except MFCC			All Feature Subsets		
	Accuracy	**F1-Score**	**MCC**	**Accuracy**	**F1-Score**	**MCC**	**Accuracy**	**F1-Score**	**MCC**
Naive Bayes	0.65	0.67	0.29	0.81	0.81	0.51	0.83	0.83	0.54
Logistic Regression	0.81	0.79	0.45	0.83	0.82	0.51	0.85	0.84	0.57
k-NN	0.82	0.79	0.48	0.84	0.82	0.53	0.85	0.82	0.56
Multilayer Perceptron	**0.83**	**0.81**	**0.50**	0.81	0.80	0.46	0.84	0.83	0.54
Random Forest	0.79	0.78	0.40	0.83	0.82	0.51	0.85	0.84	0.57
SVM (Linear)	0.81	0.80	0.46	0.84	0.83	0.54	0.83	0.82	0.52
SVM (RBF)	**0.83**	**0.81**	**0.50**	0.83	0.81	0.50	**0.86**	**0.84**	**0.59**
Average	0.79	0.77	0.43	0.83	0.82	0.51	0.84	0.83	0.55
Std. Dev.	0.07	0.05	0.08	0.01	0.01	0.03	0.01	0.01	0.02
Ensemble with Voting	0.81	0.80	0.46	**0.85**	**0.84**	**0.57**	0.85	0.84	0.58
Ensemble with Stacking	0.82	0.81	0.49	0.83	0.81	0.52	0.84	0.82	0.55

4. Discussion

In this study, we have presented a detailed analysis of signal processing techniques used in PD classification from voice recordings. The most commonly used set of features in this domain, which is referred to as "baseline features" throughout this study, has also been included as a separate group. We have collected the voice recordings of 252 subjects (188 PD patients and 64 healthy controls) in the context of this study, extracted various feature subsets from the voice recordings and evaluated the effectiveness of each feature subset and also their combination using a number of classifiers. The predictions of the individual classifiers were combined with ensemble stacking and voting approaches and comparative analysis is presented over these ensemble accuracies in order to decrease the classifier bias. We also performed feature selection on the combination of all feature subsets analyzed in this study aiming at determining the optimal minimal subset of features.

In addition to the speech signal processing techniques used in the related domain, in this study we used, to the best of our knowledge for the first time, the TQWT for feature extraction in PD classification. In our experiments, TQWT showed promising results by achieving the performance of the state-of-the-art techniques used in the context of discriminating healthy subjects from PD patients based on dysphonia measures. The mRMR filter also showed that the TQWT features carry important unique discriminative information in separating healthy subjects from PD patients. In the TQWT based feature extraction part, in addition to energy values, Shannon entropy and Log Energy entropy values, which are both used to quantify how much information is carried in the relevant subband, are calculated after decomposing speech signals into subbands. The mRMR rankings showed that the Log Energy entropy features that have been extracted from the subbands representing higher frequencies are among the most discriminative features. This implies that in PD speech samples, there is a significant change in the amount of information carried in higher frequencies and this can be related with the impact of incomplete vocal fold closure that exhibits increased aero-acoustic noise. In the calculation process of Log Energy entropy features, base-10 logarithm of the squares of wavelet coefficients are taken and this logarithmic effect unveils the importance of small changes in the high frequency information which increases the discriminative power of models.

MFCCs have produced the second-best results in our experiments. The mRMR rankings showed that MFCCs and TQWT coefficients contain complementary information that provides higher classification accuracy when used together in the PD classification problem. This situation may occur due to the frequency domain characteristics of the filter-banks used in the extraction of MFCCs and TQWT. In the MFCC, the linear frequency axis information, obtained as the output of discrete Fourier Transform, is mapped to mel scale to increase the performance of algorithm by imitating the human hearing system. During this mapping operation two types of filter-banks are employed, below 1000 Hz the filters are localized linearly and above 1000 Hz the filters spread logarithmically. On the other hand, in the TQWT, for all frequency values, constant Q-factor filters are employed with a logarithmic localization resulting in a better model of human hearing. Additionally, the frequency domain representations of filters used in the extraction of MFCCs have triangle shape while the TQWT analysis filters' frequency responses are bell-shaped with smoother transitions which may result better frequency localization to catch abnormalities in PD patients. As a final comment, unlike the MFCC extraction process in which the temporal information is lost under the employed window after applying discrete Fourier transform, the TQWT ensures temporal localization during the transform for the relevant subband. In PD speech samples, possible disruptions in vocal fold may show themselves as transient waveforms and these abnormalities can be detected with a higher success with the TQWT.

Another important contribution of this study is the comparison of the signal processing methods with different types of classifiers. We should note that combining feature subsets and selecting a minimal subset of features using mRMR feature selection approach improved the highest accuracies obtained with all the classifiers except multilayer perceptron. The highest accuracy of 0.86 with 0.84 F1-score and 0.59 MCC is achieved by feeding top-50 features selected by mRMR to SVM-RBF classifier. While more than a half of these features belongs to the TQWT feature subset, the remaining features belong to MFCC, vocal fold, bandwidth, formant and F_0 related wavelet feature subsets.

We should note that higher accuracies than the best accuracy obtained in this study have been reported in the literature for PD classification problem. However, although the speech datasets used in these studies contain multiple speech recordings per subject, most of these studies use leave-one-out cross validation technique which results in biased predictive models by sparing some samples of an individual in the training and some for the testing. This process creates an artificial overlap between the training and test sets. It has been shown that when unbiased validation is performed, the accuracies of the proposed models dramatically decrease. We remark that the main goal of this paper is to compare the feature extraction techniques used in speech-based PD classification problem with the proper unbiased techniques and also to assess the effectiveness of the TQWT technique in this problem rather than improving the accuracies of the existing models. As a future direction, the TQWT technique, which has showed promising results in PD classification problem, can be used to predict the Unified Parkinson's Disease Rating Scale (UPDRS) score of PD patients to build a robust PD telemonitoring system.

Acknowledgements

The study has the approval of Clinical Research Ethics Committee of Bahcesehir University. This research was supported by The Scientific and Technological Research Council of Turkey (TUBITAK) grant number 215E008, https://www.tubitak.gov.tr/en.

References

[1] J. Jankovic, "Parkinson's disease: clinical features and diagnosis," *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 79, no. 4, pp. 368-376, 2008.

[2] M. C. de Rijk, L. J. Launer, K. Berger, M. M. Breteler, J. F. Dartigues, M. Baldereschi, L. Fratiglioni,A. Lobo, J. Martinez-Lage,C.Trenkwalder, and A. Hofman, "Prevalence of Parkinson's disease in Europe: A collaborative study of population-based cohorts," *Neurology*, vol. 54, pp. 21–23, 2000.

[3] S. K. Van Den Eeden, C. M. Tanner, A. L. Bernstein, R. D. Fross, A. Leimpeter, D. A. Bloch, and L. M. Nelson, "Incidence of Parkinson's disease: Variation by age, gender, and Race/Ethnicity," *Am J Epidem*, vol. 157, pp. 1015-1022, 2003.

[4] A. Tsanas, M. A. Little, P. E. McSharry, and O. Lorraine. Ramig, "Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests.", *IEEE transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 884-893, 2010.

[5] C. Goetz, G. Stebbins, D. Wolff, W. DeLeeuw, H. Bronte-Stewart, R. Elble, M. Hallett, J. Nutt, L. Ramig, T. Sanger, A. Wu, P. Kraus, L. Blasucci, E. Shamim, K. Sethi, J. Spielman, K. Kubota, A. Grove, E. Dishman and C. Taylor, "Testing objective measures of motor impairment in early Parkinson's disease: Feasibility study of an at-home testing device", *Movement Disorders*, vol. 24, no. 4, pp. 551-556, 2009.

[6] B. E. Sakar, M. Isenkul, C. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin and O. Kursun, "Collection and Analysis of a Parkinson Speech Dataset With Multiple Types of Sound Recordings", *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 4, pp. 828-834, 2013.

[7] C. O. Sakar and O. Kursun, "Telediagnosis of Parkinson’s Disease Using Measurements of Dysphonia", *Journal of Medical Systems*, vol. 34, no. 4, pp. 591-599, 2009.

[8] A. Tsanas, M. Little, P. McSharry and L. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity", *Journal of the Royal Society Interface*, vol. 8, no. 59, pp. 842-855, 2010.

[9] H. Gürüler, "A novel diagnosis system for Parkinson’s disease using complex-valued artificial neural network with k-means clustering feature weighting method", *Neural Computing and Applications*, vol. 28, no. 7, pp. 1657-1666, 2016.

[10] M. A. Little, P. Mcsharry, E. Hunter, J. Spielman, and L. Ramig, "Suitability of Dysphonia Measurements for Telemonitoring of Parkinsons Disease," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1015–1022, 2009.

[11] M. Peker, "A decision support system to improve medical diagnosis using a combination of k-medoids clustering based attribute weighting and SVM", *Journal of Medical Systems*, vol. 40, no. 5, 2016.

[12] B. Erdogdu Sakar, G. Serbes and C. Sakar, "Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease", *PLOS ONE*, vol. 12, no. 8, p. e0182428, 2017.

[13] M. Peker, B. Şen and D. Delen, "Computer-Aided Diagnosis of Parkinson's Disease Using Complex-Valued Neural Networks and mRMR Feature Selection Algorithm", *Journal of Healthcare Engineering*, vol. 6, no. 3, pp. 281-302, 2015.

[14] A. Tsanas, M. Little, P. McSharry, J. Spielman and L. Ramig, "Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease", *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1264-1271, 2012.

[15] I. Selesnick, "Wavelet Transform with Tunable Q-Factor", *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 3560-3575, 2011.

[16] I. Selesnick, "Resonance-based signal decomposition: A new sparsity-enabled signal analysis method", *Signal Processing*, vol. 91, no. 12, pp. 2793-2809, 2011.

[17] I. Bayram, "An Analytic Wavelet Transform with a Flexible Time-Frequency Covering", *IEEE Transactions on Signal Processing*, vol. 61, no. 5, pp. 1131-1142, 2013.

[18] I. Bayram and Ö. Akyıldız, "Primal-dual algorithms for audio decomposition using mixed norms", *Signal, Image and Video Processing*, vol. 8, no. 1, pp. 95-110, 2013.

[19] I. Bayram and M. Kamasak, "A Simple Prior for Audio Signals", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1190-1200, 2013.

[20] H. Peng, F. Long and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, 2005.

[21] Y. Yunusova, G. Weismer, J. Westbury and M. Lindstrom, "Articulatory Movements During Vowels in Speakers with Dysarthria and Healthy Controls", *Journal of Speech Language and Hearing Research*, vol. 51, no. 3, p. 596, 2008.

[22] T. Falk, W. Chan and F. Shein, "Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility", *Speech Communication*, vol. 54, no. 5, pp. 622-631, 2012.

[23] P. Boersma, "Praat: Doing Phonetics by Computer", *Ear and Hearing*, vol. 32, no. 2, p. 266, 2011.

[24] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52-55, 2006

[25] J. I. Godino-Llorente, P. Gomez-Vilda, M. Blanco-Velasco, "Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters," *IEEE Transactions on Biomedical Engineering*, vol. 53, pp. 1943-1953, 2006.

[26] A. Tsanas, M. Little, P.E. McSharry, L.O. Ramig: "New nonlinear markers and insights into speech signal degradation for effective tracking of Parkinson"s disease symptom severity", *International Symposium on Nonlinear Theory and its Applications (NOLTA)*, pp. 457-460, Krakow, Poland, 5-8 September 2010.

[27] I. R. Titze, *Principles of Voice Production*, National Center for Voice and Speech, Iowa City, US, 2nd printing, 2000.

[28] A. Tsanas, M. Little, P.E. McSharry, L.O. Ramig: "Enhanced classical dysphonia measures and sparse regression for telemonitoring of Parkinson's disease progression", *IEEE Signal Processing Society*,

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 594-597, Dallas, Texas, US, 14-19 March 2010

[29] E. Alpaydin, *Introduction to Machine Learning*. London, U.K.: MIT Press, 2010.

[30] L. K. Hansen and P. Salamon, “Neural Network Ensembles", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 993-1001, 1990.

[31] A. Krogh and J. Vedelsby, “Neural network ensembles, cross validation, and active learning,” *In Advances in neural information processing systems*, pp. 231-238, 1995.

[32] S. Hashem, "Optimal Linear Combinations of Neural Networks", *Neural Networks*, vol. 10, no. 4, pp. 599-614, 1997.

[33] D. Opitz and J. Shavlik, “Generating Accurate and Diverse Members of a Neural Network Ensemble", D. Touretsky, M. Mozer and M. Hasselmo (Editors), *Advances in Neural Information Processing Systems*, vol. 8, pp. 535-54, The MIT Press, 1996.

[34] C. Sakar, O. Kursun and F. Gurgen, "A feature selection method based on kernel canonical correlation analysis and the minimum Redundancy–Maximum Relevance filter method", *Expert Systems with Applications*, vol. 39, no. 3, pp. 3432-3437, 2012.

[35] C. Ding and H. Peng, “Minimum redundancy feature selection from microarray gene expression data,” *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003.*

[36] C. O. Sakar, O. Kursun, H. Seker and F. Gurgen, "Combining multiple clusterings for protein structure prediction", *International Journal of Data Mining and Bioinformatics*, vol. 10, no. 2, p. 162, 2014.

[37] M. Pal and G. Foody, "Feature Selection for Classification of Hyperspectral Data by SVM", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 5, pp. 2297-2307, 2010.

[38] A. Idris, A. Iftikhar and Z. Rehman, "Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling", *Cluster Computing*, pp. 1-15, 2017.

[39] Moro-Velázquez, L., Gómez-García, J.A., Godino-Llorente, J.I., Villalba, J., Orozco-Arroyave, J.R. and Dehak, N. “Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's Disease”, *Applied Soft Computing*, 62, pp. 649-666, 2018.