Medical & Biological Engineering & Computing (2020) 58:2757–2773
https://doi.org/10.1007/s11517-020-02250-5

ORIGINAL ARTICLE

Estimation of Parkinson's disease severity using speech features and extreme gradient boosting

Hunkar C. Tunc[1,2] · C. Okan Sakar[1] · Hulya Apaydin[3] · Gorkem Serbes[4] · Aysegul Gunduz[3] · Melih Tutuncu[3] · Fikret Gurgen[5]

Received: 28 January 2020 / Accepted: 20 August 2020 / Published online: 10 September 2020

Abstract

In recent years, there is an increasing interest in building e-health systems. The systems built to deliver the health services with the use of internet and communication technologies aim to reduce the costs arising from outpatient visits of patients. Some of the related recent studies propose machine learning–based telediagnosis and telemonitoring systems for Parkinson's disease (PD). Motivated from the studies showing the potential of speech disorders in PD telemonitoring systems, in this study, we aim to estimate the severity of PD from voice recordings of the patients using motor Unified Parkinson's Disease Rating Scale (UPDRS) as the evaluation metric. For this purpose, we apply various speech processing algorithms to the voice signals of the patients and then use these features as input to a two-stage estimation model. The first step is to apply a wrapper-based feature selection algorithm, called Boruta, and select the most informative speech features. The second step is to feed the selected set of features to a decision tree–based boosting algorithm, extreme gradient boosting, which has been recently applied successfully in many machine learning tasks due to its generalization ability and speed. The feature selection analysis showed that the vibration pattern of the vocal fold is an important indicator of PD severity. Besides, we also investigate the effectiveness of using age and years passed since diagnosis as covariates together with speech features. The lowest mean absolute error with 3.87 was obtained by combining these covariates and speech features with prediction level fusion.

Keywords Unified Parkinson's Disease Rating Scale · UPDRS prediction · Machine learning · Telemonitoring · E-health

1 Introduction

Parkinson's disease (PD) is a neurological disorder characterized by various motor (e.g., tremor, speech impairment, postural instability) and non-motor (e.g., constipation, cognitive impairment, depression, sleep disorders) symptoms [31]. It is estimated that as of 2016, more than 6 million people worldwide are living with PD [85]. Aging is considered to be the biggest risk factor for the development and progression of PD [28]; therefore, the number of people with PD condition is expected to rise with the aging population [13, 84]. There is still no definitive test for the diagnosis and evaluating the severity of PD [38]. Diagnosis and symptom monitoring typically require patients to visit a medical clinic. It was estimated that the economic burden caused by PD was over $14.4 billion in 2010 in the USA alone [32].

Recently, there are some research efforts that aim to diagnose and monitor PD using machine learning techniques [1, 2, 5, 47, 65, 83]. The main goal of these studies is to develop a telemedicine system that improves the access of PD patients to medical services. These studies are mostly based on the assessment of the severity of various PD symptoms in an invasive way. The literature studies show that speech disorders are one of the most suitable PD symptoms that can be used in such PD telediagnosis and telemonitoring studies [3, 63, 78]. This has two main reasons. Firstly, speech disorders are seen in approximately 90% of the PD patients [77]. And secondly, speech disorders are one of the earliest signs of PD [25].

The estimation of the UPDRS scores from speech tests was first conducted by Tsanas et al. [77] and followed by many efforts that are summarized in Section 2. Motivated from the studies showing the potential of speech disorders in PD telediagnosis and telemonitoring systems, in this

✉ Hunkar C. Tunc
huenkar.tunc@uni-konstanz.de

Extended author information available on the last page of the article.

study, we aim to estimate the Unified Parkinson's Disease Rating Scale (UPDRS) subscores of PD patients from their voice recordings. UPDRS [16] is a well-accepted standard for evaluating the severity of PD and the most commonly used scale in clinical trials [59, 71, 87]. Specifically, in this paper, we focus on improving the generalization ability of the existing dysphonia-based UPDRS estimation models by using extreme gradient boosting (XGBoost) algorithm, whose success has been shown in many recent studies [10], in a fully unbiased way by preventing the issues that may lead to over-optimistic results. Another contribution of our study is to use the tunable-Q wavelet transform (TQWT) technique to extract features from the speech signals of the subjects. The success of TQWT coefficients in speech-based PD classification systems has recently been shown [65]. Based on this finding, in this study, we use TQWT coefficients in the UPDRS estimation model along with the other most effective signal processing techniques.

Studies reported that PD symptoms tend to progress linearly [9] and aging is the most important factor in symptom progression [28]. Due to this, in this study, we investigate the contribution of the predictive information obtained from a set of covariates, which are age, years since PD diagnosis, and gender of the subject, to the speech-based predictive model. For this purpose, we combined the covariates with speech features by data level and prediction level fusion methods and show to what extent the covariates help to decrease the UPDRS estimation error of the speech-based system.

The proposed system consists of two main steps. First, we have applied time, frequency, time-frequency, and time-scale based linear/non-linear signal processing algorithms to the vowel /a/ samples obtained from the PD patients with the aim of extracting the characteristics of the malfunctions (called as features) in the voice produced by Parkinson's patients. Later, these extracted features were given to machine learning methods to map the results of signal processing algorithms to motor-UPDRS and total-UPDRS scores. The curse-of-dimensionality problem, which arises when the number of features exceeds the number of samples in a machine learning problem, is addressed with a wrapper-based feature selection algorithm, Boruta [34].

The datasets used in this domain typically contain multiple recordings per subject [63, 77, 78] and an important issue that should be taken into account while building machine learning models on such datasets is the use of a suitable cross-validation (CV) strategy. Otherwise, CV techniques such as the simple 10-fold CV may lead to optimistic results since the training and test sets may contain different recordings of the same subject. Many studies have addressed this issue and suggested the use of subject-wise CV methods in the presence of this type of data [5, 47, 62–64, 83]. Based on these findings, we apply a subject-wise cross-validation strategy throughout our experiments. Feature subset selection bias, which occurs when all the samples are used in the selection of a subset of features [17, 45, 73], should also be avoided to construct realistic diagnosis and monitoring systems. We also avoid feature selection bias by applying the feature selection algorithm on the training set only and finally use the selected features on the left-out test set.

We applied the proposed UPDRS prediction system on a dataset collected in the context of our project. The dataset consists of 305 recordings belonging to 86 PD patients. We have used this dataset before for PD classification in [65] and made classification dataset publicly available in UCI Machine Learning Repository [4]. Besides, we applied the proposed system to another public dataset and presented the UPDRS prediction results in comparison with another related work that used an unbiased cross-validation procedure.

The remaining of this paper is organized as follows. Section 2 gives the previous UPDRS estimation studies that use speech signals with machine learning models. Section 3 presents the materials and methods including the description of the dataset, feature extraction and selection techniques, cross-validation strategy, and statistical tests used throughout this study. Section 4 includes the results of the experiments. Finally, we give conclusions and discussions in Section 5.

2 Related works

There is increased attention in investigating the potential of telemonitoring technologies for improving the quality of the severity assessment and symptom monitoring of PD [15, 74]. It has been well-studied in the literature that people with Parkinson's disease commonly suffer from speech disorders [26, 27, 29, 37, 61, 89]. Also, speech disorders may be seen as one of the earliest signs of PD [14, 25]. There is supporting evidence that speech degradation symptoms, such as *hypophonia* (reduction in voice amplitude) and *dysphonia* (increased breathiness and hoarseness in the voice), occur with the progress in PD [29, 50]. Therefore, speech tests, which may consist of running speech and/or sustained vowel phonations, are employed to assess the degree of speech degradation. Commonly, sustained vowels, in which the subject is requested to hold the frequency and amplitude of phonation constant as long as possible, are used in PD clinical in order to avoid some confounding effects of articulatory movement in running speech [66].

There are three main parts in the UPDRS: Section I is referred as "Mentation, Behavior and Mood," Section II is referred to as "Activities of Daily Living," and Section III is referred to as "Motor Examination." We refer to Section III of the UPDRS as motor-UPDRS, and combination of all sections as total-UPDRS. In literature, there are some studies attempted to find a mapping between dysphonia measures and UPRDS. In one of these studies, Midi et al. [44] examined the relation between various voice parameters with motor-UPDRS on a test group of 20 early-stage PD patients, none of which was suffering from speech problems. The authors reported only several significant correlations, and they discuss that the choice of subject group might have an influence on the results. In a later study, Majdinasab et al. [39] investigated the relation between motor-UPDRS and Voice Handicap Index (VHI). VHI is a self-assessed test which measures the effect of voice disorders on daily life [30]. Majdinasab et al. [39] selected patients who were suffering from PD for at least five years and proclaimed a voice disorder related to PD. They reported a positive correlation between VHI and motor-UPDRS ($r = 0.485$, p value $< .05$).

The study by Goetz et al. [23] analyzed the practicality of computer-based at-home testing device in monitoring early-stage PD patients. They recruited fifty-two PD patients and monitored them for a 6-month period. Fifty patients completed the trial and forty-eight of the patients remained unmedicated during the study. Patients performed a number of motor tasks, including speech tests, on a weekly basis. UPDRS assessments were conducted at the beginning, after 3 months, and at the end of the trial. The estimation of the UPDRS from speech tests was first conducted by Tsanas et al. [77] on this dataset. They performed piecewise linear interpolation in order to generate the weekly UPDRS estimates. Their basis for applying linear interpolation was that PD tends to progress linearly [9], especially in unmedicated patients [67]. They estimated the UPDRS scores with a mean absolute error (MAE) of 5.95±0.19 for motor-UPDRS and 7.52±0.25 for total-UPDRS using simple 10-fold cross-validation (CV) and showed the potential of the use of machine learning algorithms based on dysphonia measurements in the monitoring of PD. A later study [78], by using the same CV method and data used in [77], reported MAE of approximately 2 UPDRS points (p value $< .001$) for both motor-UPDRS and total-UPDRS.

Naranjo et al. [47] used the dataset made available by Tsanas et al. [77]. They proposed a Bayesian linear regression method for handling the replicated measurements and also the time factor. As mentioned above, the UDPRS values between two real UPDRS measurements in this dataset were filled with linear interpolation and the authors used only the real UPDRS measurements in their study. They obtained MAE of 7.52±1.10 for motor-UPDRS and 9.64±1.64 for total-UPDRS and claimed that their results are the first reliable ones published on this dataset.

Bayestehtashk et al. [5] formed another dataset for predicting the UPDRS from speech which consists of 168 PD patients. They conducted three different speech tasks: sustained phonation task, diadochokinetic task, and a reading task. They used a subject-wise CV and obtained a MAE of 5.5 for predicting the motor-UPDRS by combining the features extracted from all three tasks.

Zhan et al. [90] generated a mobile Parkinson's disease score (mPDS) for assessing the PD severity by using the sensor data captured from smartphones. There are five tasks that determine the mPDS measure: voice, finger tapping, gait, balance, and reaction time. They measured the correlation of mPDS with UPDRS on a test group of 40 participants (23 with PD and 17 healthy subjects). They reported strong correlations between mPDS and total-UPDRS ($r = 0.81$, p value $< .001$) and motor-UPDRS ($r = 0.88$, p value $< .001$). However, the authors point out that the participants of their study were not representing the general population in PD. Most of the subjects were college graduates and familiar with smartphones.

Buza and Varga [8] used feedforward neural networks to predict the UPDRS scores on the datasets in [63, 77] by also taking into consideration the hubness effect. They only considered a subclass of features, namely jitter and shimmer features available in the datasets and used a subject-wise cross-validation for the evaluation of the results. They proposed a hubness-aware error-correction method and demonstrated that the proposed approach improves the prediction accuracy.

One of the recent studies in this context [48] focused on predicting the depression element, which is a specific part of UPDRS, from voice recordings of PD patients. The study relied on patient self-assessments of existence or absence of depression and achieved the highest accuracy of 0.77 by using random forest algorithm.

3 Material and methods

3.1 Dataset description

The first dataset used in this study was collected in the context of our project at the Department of Neurology in Cerrahpasa Medical Faculty, Istanbul University-Cerrahpasa, and has been used for PD classification in our recent work [4, 65]. The UPDRS score is available for 86 PD patients (49 male and 37 female) out of 188 who participated in the study, and their summary information is demonstrated in Table 1. Seven patients participated in the study more than once at different times, and a total of 93 sessions were held. Subjects were requested to produce a constant phonation of the vowel /a/ for at least 5 s. At each session, subjects performed this procedure for three to five successful times and in a total of 305 sound recordings were acquired. The speech data was recorded in a silent room with a 44.1-kHz

Table 1 Summary information of the test group

Variables	Males (n = 49)	Females (n = 37)	All subjects (n = 86)
Age	64.69±9.60	64.43±9.88	64.58±9.66
Years since PD diagnosis	5.12±4.23	6.13±5.31	5.55±4.72
motor-UPDRS	12.99±4.72	11.38±5.21	12.29±5.21
total-UPDRS	20.90±9.04	20.09±11.22	20.54±10.05

Statistics are given in the form mean ± standard deviation

sampling rate and 16-bit resolution settings. All participants were informed about the study and gave written consent. No subject exclusion criteria were applied.

The total-UPDRS and motor-UPDRS scores range from 0 to 124 and 0 to 56, respectively, in the UPDRS scale applied to the patients in the context of this study. Due to different practices in the calculation of the UPDRS score, the UPDRS scores in some other studies discussed in the introduction section range from 0 to 108 for motor-UPDRS and 0 to 176 for total-UPDRS. In both practices, 0 corresponds to a healthy state and the maximum value corresponds to the highest severity of symptoms.

In addition to the dataset described above, we also utilized the publicly available Parkinson's telemonitoring dataset [77] in order to be able to compare our approach with the existing literature that reported results on this dataset. This dataset consists of approximately 6000 recordings of the sustained vowel /a/ coming from 42 early-stage PD patients. These patients were monitored for a 6-month period, and the voice recordings were obtained on a weekly basis. UPDRS assessments were conducted at the beginning, after 3 months, and at the end of the trial, and piecewise linear interpolation was performed in order to generate the weekly UPDRS estimates. The total-UPDRS and motor-UPDRS scores in this dataset range from 0 to 176 and 0 to 108, respectively. The raw voice recordings are not publicly available; however, the data analyzed in [77] with the extracted features is publicly available at the UCI repository [4]. Parkinson's telemonitoring dataset is a widely used dataset in the context of PD telemonitoring studies based on speech recordings. Further details about the dataset can be found in [77].

3.2 Feature extraction

In the production of healthy vowel voices, successive opening and closure movements of the vocal fold produce nearly periodic signals called phonemes. During this process, the time durations between various vocal fold positions, in which they are apart or in collision, remain almost the same for successive cycles. The reciprocal of this duration is generally named as the fundamental frequency (F_o) of the phoneme. In PD patients, the usual vibration pattern of the vocal folds is severely affected due to the decrease occurred in the amount of dopamine, which is an essential neurotransmitter used in neuron communication. Besides, a high-energy turbulent noise component also occurs due to the incomplete closure of vocal fold resulting in dramatic changes in the power distribution of speech signal [21, 65].

In this study, linear and non-linear signal processing algorithms, based on time, frequency, and time-frequency and time-scale representations, were applied to recorded /a/ vowels in order to quantitatively unveil the impairments seen in PD patient's speech behaviors. The measures obtained with these algorithms are called dysphonia measures. As the first part of the extracted features, traditional signal processing algorithms were applied to collected speech samples with the aim of obtaining linear behavior time, frequency, and time-frequency based signal features which include jitter, shimmer, fundamental frequency (F_0), harmonics-to-noise ratio (HNR), noise-to-harmonics ratio (NHR), intensity, formant frequencies, bandwidths [42, 86], root mean square (RMS) energy, strength of excitation (SoE) [46], and cepstral peak prominence (CPP) measures. The details of these features given in Table 2 are as follows. As the representative of F_0 measures, the standard deviation, median, minimum, mean, and maximum values of F_0 were employed. Similarly, minimum, maximum, and mean intensity values were used as features. As the measure of tongue movement, the first four formants were employed as features. The first four bandwidths of the formants were used as bandwidth features which are mostly related to volumes of vocal tract cavities. The CPP feature represents the degree of regularity or periodicity in the voice signal. Higher CCP values correspond to greater periodicity.

After extracting the traditional baseline measures, three additional signal processing approaches, named as recurrence period density entropy (RPDE), detrended fluctuation analysis (DFA), and pitch period entropy (PPE), were applied to the collected speech samples [77]. The DFA characterizes the turbulent noise component that is caused by air-flow in the vocal tract. RPDE quantifies the

Table 2 Features obtained with traditional signal processing algorithms

Measure	Toolbox	Explanation
Jitter Variants	Praat and VAT	Jitter variants quantify the cycle-to-cycle variations in F_0.
Shimmer variants	Praat and VAT	Shimmer variants quantify the cycle-to-cycle changes in the amplitude of signal of interest.
F_0 measures	Praat and VAT and VS	F_0 is the frequency of vocal fold opening and closure movements.
Harmonics to noise ratio and noise to harmonics ratio	Praat and VS	Represent the ratio of signal information power and the noise power, which occurs due to the incomplete vocal fold closure.
Intensity parameters	Praat	Quantify the power of speech signal in dB.
Formant frequencies	Praat and VS	Specific frequency lines that are strengthened by the vocal tract.
Bandwidth	Praat	Metric that quantifies the frequency range between the formant frequencies.
Root mean square (RMS) energy	VS	Quantifies the RMS energy of speech at every frame controlled by a variable window.
Strength of Excitation (SoE)	VS	Quantifies the relative amplitude of impulse-like excitation derived from the instant of significant excitation of the vocal-tract system [46].
Cepstral peak prominence (CPP)	VS	Measures the relative amplitude of the cepstral peak prominence with regards to the expected amplitude obtained with linear regression.

VAT voice analysis toolbox, *VS* voice sauce deviations from the periodicity of vocal fold oscillations. In PD speech samples, it is challenging to sustain stable F_0 values due to the incomplete vocal fold closure. In this respect, PPE quantifies the impaired control of the fundamental frequency by using a logarithmic scale. Additional to these vocal fold movement–related features, as an alternative, empirical mode decomposition (EMD) based features were also calculated. By using EMD, the speech samples were decomposed into a small number of intrinsic mode functions (IMF) which form a complete and nearly orthogonal basis for the PD speech samples. The first few IMFs of the signal of interest represent the high-frequency components and these high-frequency bearing IMFs were given to Teager-Kaiser energy, squared energy, and Shannon entropy operators to characterize the noise. The details of these features can be found in Table 3.

The wavelet transform (WT) is a powerful tool for processing non-stationary signals due to its adaptive time-scale representation property which provides a good frequency resolution at low frequencies and a good time resolution at high-frequency signal components. Due to its flexible time-frequency representation ability, WT has previously been applied to various biomedical signals [33, 52, 53, 69, 82] including the speech signals taken from PD patients [41, 65].

The motivation behind the use of WT in speech-based PD applications is that a healthy subject is expected to be able to sustain a stable pitch for a vowel while significant deviations would occur in the fundamental frequency of the speech signals produced by the PD patients. In this respect, two various approaches have been carried out in the literature. In the first approach, the fundamental frequency (F_0) of the speech signals is obtained in the first sense and later the discrete WT (DWT), which is a fast implementation of ordinary WT based on filter-bank theory, is applied to F_0 signal in order to quantify the deviations from the expected behavior [80]. In our study, we applied 10 levels of DWT to F_0 and the resultant detail/approximation coefficients were given to energy, Shannon's entropy, the log energy entropy, and the Teager–Kaiser energy operators resulting in feature subsets related with the fundamental frequency of vowels. This feature subset is referred to as the "WT features related with F_0" in Table 3. As the second WT-based approach, the tunable Q-factor wavelet transform (TQWT) [68], in which the Q-factor of the decomposition/reconstruction filters can be tuned according to the time-frequency behavior of processed signal, was applied to the recorded raw speech signals with the aim of extracting adjusted time-scale features. Q-factor of a decomposition/reconstruction filter can be defined as the ratio of the center frequency of that filter to its bandwidth. By definition, when a better frequency resolution is needed (for example, in order to analyze the sustained oscillations seen in vowels), relatively high Q-factor filters can be chosen. On the other hand, relatively low Q-factor filters can be preferred for analyzing transient signals such as the deviations that can be occurred in the expected successive opening and closure movements of the vocal fold. Hence, in the proposed study, the TQWT is employed to form the optimum time-frequency representation that unveils the effects of PD in collected speech samples. In the TQWT, three parameters named as J (the number of decomposition and reconstruction levels), Q (Q-factor of band-pass filters), and r (redundancy or oversampling rate) are used to tune the applied WT according to the signal of interest. In the proposed study, after the intuitive testing phase, the optimum parameter

Table 3 Other features extracted from speech samples

Measure	Toolbox	Explanation
Recurrence period density entropy (RPDE)	VAT	Measures the level of uncertainty in the quantification of fundamental frequency.
Detrended fluctuation analysis (DFA)	VAT	Quantifies the degree of stochastic self-similarity in the turbulent noise.
Pitch period entropy (PPE)	VAT	Quantifies the degree of disruption in fundamental frequency using a logarithmic scale.
Mel frequency cepstral coefficients (MFCCs)	VAT	Quantifies the degree of impairments caused by PD in vocal tract separately from the vocal folds.
Wavelet transform (WT) features related with F_0	VAT	Employed to reveal the changes in F_0 with PD progression.
Glottis quotient (GQ)	VAT	Measures the opening and closing time-intervals of the glottis.
Glottal to noise excitation (GNE)	VAT	Measures the extent of turbulent noise that is caused by incomplete vocal fold closure.
Vocal fold excitation ratio (VFER)	VAT	Aims to explain the nonlinear physiological phenomena in speech production.
Empirical mode decomposition (EMD)	VAT	Decomposes a speech signal into elementary signal components called as intrinsic mode functions (IMFs).
TQWT-based features	TQWTT	Has the ability to represent PD speech samples in an adjustable scale.

TQWTT tunable Q-factor wavelet transform toolbox set giving the best representation was found as $J=35$, $Q=2$, and $r=4$. A relatively high Q-value was chosen to process speech signals which mostly consist of sustained oscillations. After extracting the detail and approximation coefficients, the energy/entropy values of each decomposed level were obtained, and these energy/entropy values were used in UPDRS prediction. The details of the applied TQWT procedure can be found in [65] for PD speech samples with more detail.

In literature, mostly, linear signal processing methods have been employed to extract useful information from the PD speech samples [61]. Although these linear speech analysis methods have performed well in modelling the deviations that occurred in the behavior of PD samples, new nonlinear speech processing methods have been started to be used more often in representing the characteristics of PD speech samples due to their higher modelling capability [78]. In this respect, three non-linear speech processing methods named as the glottal to noise excitation (GNE), vocal fold excitation ratio (VFER), and glottis quotient (GQ) were utilized in more recent studies for extracting additional information from PD speech samples. The GNE parameter gives information about whether the analyzed speech signal originates from vibrations of the vocal folds or from the acoustic noise generated in the vocal tract [43]. The VFER metric works on the vocal fold cycles in order to measure the energy ratios during each cycle. The main aim is to measure the nonlinear physiological phenomena in speech production which occur as a result of incomplete vocal fold closure. Lastly, the GQ parameter quantifies the opening and closing duration of the glottis.

Mel-frequency cepstral coefficients (MFCCs), which imitate the efficient filtering abilities of the human ear, are widely used as a robust feature extractor in the field of speech processing [20, 40, 81]. During the MFCC-based feature extraction, cepstral analysis is combined with spectral domain partitioning by employing overlapped filter banks that have triangular shape frequency responses resulting in a dense representation of the spectrum. For PD studies, the advantage of using MFCC comes from their ability to detect subtle changes in the motion of the articulators (tongue, lips) which are known to be severely affected in PD patients [79]. In the proposed study, the MFCC method was applied to recorded signals with the aim of extracting MFCC matrices containing segmental information (MFCC coefficients). After obtaining the MFCC coefficients, the mean, standard deviation, log-energy, and the first/second derivatives of them were calculated and later given as features into learning models. In the feature extraction part, all extracted features were obtained by using Praat [7], voice analysis toolbox [76, 78, 80], voice sauce [70], and tunable Q-factor wavelet transform (TQWT) [68] software packages.

3.3 Feature selection

We extracted 872 features (p) from 305 samples (N) belonging to 86 PD patients. In this case, p could be considered as much larger than N, often denoted as $p \gg N$. The major issues that should be dealt with this type of settings in machine learning problems are the high variance of the fitted model and overfitting [17]. Therefore, we applied feature selection as a pre-processing step in order to reduce the dimensionality of the dataset.

Feature selection is a common practice for aiming to enhance the predictive performance, obtaining faster models and having a better understanding of high-dimensional datasets [24]. We applied Boruta feature selection algorithm [35] due to its successful applications in various domains [51, 55, 56, 88]. Boruta is a heuristic algorithm which aims to find all relevant variables. It utilizes a random forest algorithm in the process of searching for all relevant features. For each feature in the dataset, the Boruta algorithm creates a matching feature by shuffling the values of the original feature. Random forest algorithm computes a variable importance score for every feature and then a z-score is calculated by dividing the variable importance score with its standard deviation. Boruta algorithm determines the significance of a feature by comparing its z-score with the maximum z-score among the randomized features. For more details about the Boruta feature selection algorithm, we refer the reader to [35].

3.4 Statistical analysis

We used the XGBoost [10] implementation of the gradient boosting decision tree (GBDT) [18, 19] for building the predictive models. GBDT is a machine learning algorithm successfully used in various domains, including energy, transportation, and medical [11, 75, 91]. GBDT is essentially a tree ensemble method where a large number of decision trees are combined together to build a robust model. A decision tree is a predictive model where a mapping from the observations to the target value is achieved by arranging a set of rules in a tree-like structure with respect to some given loss function. The rules correspond to the threshold values for which the feature variables are split into two disjoint sets. The decision tree grows by continuously partitioning the existing leaf nodes until some stopping criteria are satisfied. The combination process of the decision trees is conducted by a procedure known as gradient boosting where each new decision tree is constructed by fitting the residuals of the previous trees. We performed a random search [6] at each iteration on the training set for the hyper-parameter tuning of GBDTs.

Studies reported that PD symptoms tend to progress linearly [9] and aging is the most important factor in symptom progression [28]. Therefore, we used age, years since PD diagnosis, and also gender as the covariates. We combined the covariates with speech features by data level and prediction level fusion methods. For data-level fusion, we merged the covariates with the speech features and trained our model on the obtained single dataset. For prediction level fusion, we fitted a linear regression model on the covariates and then combined the predictions of the covariates and speech features by averaging. Rahn et al. [58] observed some differences in vocal pathologies of male and female PD patients. Based on this finding, we also included gender information along with the speech features.

3.5 Cross-validation

Cross-validation (CV) is a standard method for estimating the prediction error of a model by splitting the data into training and test sets [17]. In our dataset, each subject has more than one recording. If we randomly split the data into training and test sets, the recordings coming from the same subject may occur both in the training and test sets. Thus, we used the leave-one-subject-out (LOSO) CV which prevents the records coming from the same subject from appearing in both training and test sets.

LOSO CV also resembles the use-case scenario of the model [62]. The cross-validation methodology applied in our experiments is as follows: (1) We leave the samples of one subject for testing; (2) the remaining records are divided into two parts, 75% for training and 25% for validation; (3) the model is trained on the training set with feature selection followed by different hyper-parameter values of the training algorithm; (4) optimum values of hyper-parameters and selected features are determined on the validation set; (5) the optimum model is applied on the test set. This procedure is repeated for each left-out-subject, and the average results are presented. We should note that all subjects have multiple recordings per session in our dataset, so for obtaining the final prediction, we took the average of predictions of all samples of a subject in a session. The performances of the models were evaluated by the mean absolute error (MAE) and Spearman correlation:

$$MAE = \frac{1}{n}\sum_{j=1}^{n} |y_j - \hat{y}_j| \tag{1}$$

where j denotes the index of the subject, y_j is the predicted UPDRS, $\hat{y}_j$ is the actual UPDRS, and n is the number of subjects.

4 Experimental results

4.1 Data exploration

It is seen from Table 4 that there are seven features belonging to the MFCC subset in the first eleven features which have high correlation with UPDRS values. MFCCs, which are mostly based on the human hearing perception, parse the frequency range linearly for the frequencies below 1000 Hz while the remaining higher frequency range is spaced logarithmically. In MFCC, the normal cepstral coefficients do not capture the signal energy in these various frequency ranges and they also assume that the analyzed signal is stationary, which means they do not have information related to the dynamics between time frames. Hence, the logarithmic energy values of cepstral coefficients are obtained to compensate for the former drawback, while the first/second derivative operators are applied to overcome the latter. At the end of this process, the cepstral coefficients, their logarithmic values, and the first/second derivatives form a matrix which consists of the mel-spectrum information of various time frames. Later, the mean and standard deviation of these time-related values are calculated to be employed as features in further processing steps.

According to the source-filter theory [12], the voiced speech (such as the /a/ vowel used in this study) is sourced by the vibration of the vocal fold which actually gives a response to the airflow coming from the lungs. Subsequently, the output of this periodic vibrations is processed by the vocal tract by changing its spectral shape and resulting in the voiced speech. The interaction between

Table 4 Correlation coefficients of extracted features with motor-UPDRS

Features	Correlation with motor-UPDRS
$VFER_{mean}$	−0.2846
$delta\ log\ energy_{std}$	−0.2585
$6_{th}\ delta\ MFCC\ coefficient_{std}$	−0.2576
$VFER_{SNR\_SEO}$	−0.2428
$HNR_{0-2500Hz}$	−0.2456
$VFER_{entropy}$	−0.2377
$0_{th}\ delta\ MFCC\ coefficient_{std}$	−0.2344
$3_{rd}\ MFCC\ coefficient_{mean}$	0.2334
$1_{st}\ MFCC\ coefficient_{mean}$	−0.2295
$3_{rd}\ MFCC\ coefficient\ delta_{std}$	−0.2291
$0_{th}\ delta\ delta\ MFCC\ coefficient_{std}$	−0.2273

VFER vocal fold excitation ratio, *HNR* harmonics to noise ratio, *MFCC* mel-frequency cepstral coefficient features have the highest correlation with motor-UPDRS in our dataset the vocal fold and vocal tract can be modelled as the product of their transfer functions in the frequency domain resulting in the frequency domain representation of produced speech. In this respect, it can be thought that the individual effects of vocal fold and vocal tract (in the frequency domain) can be achieved by applying a logarithm operator, which converts multiplication to addition, to the MFCCs.

In Table 4, the standard deviation of delta log energy shows the highest negative correlation with UPDRS values in terms of MFCCs. This feature gives the logarithmic energy for the first derivative of MFCCs which actually describes the logarithmic energy changes in speech fluency. As expected, we observe a high negative correlation between the delta-log-energy$_{\text{std}}$ and UPRDS; because the speed and amplitude of utterance decrease when the condition of the disease goes worse and the speech pattern becomes more monotone in high UPDRS patients. Moreover, when the other MFCC-based features are investigated, it was observed that they have also high negative correlation with UPDRS, except the mean value of the third coefficient. Regarding the standard deviation of the first derivatives for zeroth, third, and sixth MFCCs, we observed high negative correlations as well with UPDRS. This shows that a significant reduction in the temporal variability between the consecutive speech frames occurs when the UPDRS goes higher in patients. A similar pattern was also seen in the standard deviation of the second derivative of zeroth MFCC, and all these temporal variability reductions can be accepted as the result of slow and monotone speech pattern seen in high UPDRS PD patients. With respect to the mean values of the first and third MFCCs, we observed a contrasting pattern: the first MFCC experiences a negative correlation, while the third MFCC has opposite behavior. This behavior can be explained by using the source-filter theory, in which the vibrations are sourced by the movement of the vocal fold. The air flow occurring as a result of these vibrations is filtered by the vocal tract, and the voices are formed. The main frequency of this source wave has different values in males (around 100 Hz), females (around 200 Hz), and children (200–300 Hz) [54, 72]. When the frequency range that is covered by the third MFCC is investigated, it is seen that the frequency values of this source wave overlap with the third coefficient's range and this shows that the deterioration in the vocal fold caused by PD results in a positive correlation with UPDRS.

In addition to the MFCC features, three VFER family–related features also have great correlation with UPDRS values in PD patients as depicted in Table 4. The VFER is a kind of dysphonia measure which has similar conceptual justification to GNE: glottal cycles (opening and closing cycle of the vocal fold) cause a synchronous excitation of different frequency bands in speech while the turbulent noise leads to uncorrelated excitation [76]. The main idea of VFER feature family is measuring the nonlinear interactions in speech production which may happen as a result of PD. In PD patients, the healthy vibration pattern usually suffers from the incomplete vocal fold closure which leads to the creation of unwanted turbulent noise. In the calculation process of VFER, firstly the opening/closure timestamps of the vocal fold are identified to extract each glottal cycle. Secondly, the frequency spectrum of each speech cycle is divided into segments having 500-Hz shifts. Thirdly, the Hilbert envelope of these segments, which correspond to various frequency bands, are calculated. Fourthly, the cross-correlations of these pair-wise envelops are found and the maximum value among the correlations between pairs of the frequency bands are chosen. Later, the maximum value of step four is chosen as the GNE for the relevant glottal cycle. Finally, the mean, standard deviation, and entropy values of all cycle-related GNE values are named as one of the members of the VFER family. Additional to these features, Teager-Kaiser energy operator (TKEO) and squared energy operator (SEO) are applied to time-domain outputs of filtered glottal cycles, which inherit only the information of specific frequency band, for obtaining the signal to noise and noise to signal ratios for each speech sample. As seen in Table 4, the mean and entropy values of VFER have high negative correlation with UPDRS. This relation is parallel to the expected scenario because, in the high UPDRS patients, the severity of the corruption in the vocal fold vibration increases and this leads to high turbulent noise. As a result, this high turbulent noise decreases the cross-correlation values found with Hilbert envelops and the chosen maximum values are getting smaller with the increasing UPDRS values. A similar pattern is also seen in the signal to noise ratio feature of VFER family when the SEO operator is used; the turbulent noise level in the time domain cycle signals rises in the high UPDRS patients and this effect reduces the signal to noise ratio.

As a final point in Table 4, a similar effect of turbulent noise is observed on the harmonics to noise ratio (HNR) feature. Due to the impairments in vocal fold closure pattern, the turbulent noise energy increases with respect to main harmonics and this reduces the HNR ratio causing a high negative correlation with UPDRS.

4.2 Motor-UPDRS estimation

As mentioned in the Introduction section of this paper, aging is one of the most important factors in PD progress and PD symptoms tend to progress linearly with elapsed time since PD diagnosis. The correlations of age and years since PD diagnosis variables with motor-UPDRS are shown in Table 5. As seen in Tables 4 and 5, years since diagnosis is the most correlated variable among all features used in

Table 5 Correlation coefficients of age and years since PD diagnosis with motor-UPDRS

Variables	Correlation of the covariates with motor-UPDRS
Years since PD diagnosis	0.3574
Age	0.2431

our study. Age is also the most correlated sixth variable. Figure 1 shows the scaled values of these two covariates grouped by patients' motor-UPDRS scores. We also see that the motor-UPDRS scores, i.e., the severity of the PD symptoms, increase with age and years since diagnosis. Therefore, we include the related factors as additional covariates into our estimation model and aim to investigate the contribution of these covariates into the dysphonia-based estimation model.

The first step of the proposed system is to apply feature extraction techniques detailed in Section 3.2 to the raw speech recordings of the patients. Since the number of features, 802, is higher than the total number of samples, 305, we first apply the Boruta feature selection algorithm and then feed the selected features to XGBoost for UPDRS estimation. Figure 2 shows the number of selected features from each speech signal processing tool used in this study. It also shows the proportion of selected features to the total number of features for each speech processing tool. It is seen that almost half of the features extracted with voice sauce toolbox (VAT) have been selected by Boruta. As seen in Tables 2 and 3, this toolbox mostly gives features that emphasize the relation between turbulent noise energy and the energy of the main source voice plus its harmonics. This situation indicates that the vibration pattern of the vocal fold, which acts as the direct source of speech harmonics and also is the reason for the turbulent noise happening due to the impairments in its own closure pattern, is highly affected in PD patients. We also see in Fig. 3 that almost 15% of the tunable-Q wavelet transform (TQWT) coefficients are selected by Boruta showing that TQWT features are effective in catching the changes in speech due to PD progression.

Table 6 shows the mean absolute error (MAE) and Spearman correlation of the proposed approach. We see that the best results with MAE of 3.87 and Spearman correlation of 0.46 are obtained by combining speech features and covariates using prediction level fusion. We should note that the covariates when used alone are fed to linear regression, while they are fed to XGBoost when fused with speech features with data-level fusion. This is due to the fact that there are only three variables in the covariates feature set and the relations among these features and also their relations with motor-UPDRS are not too complex to be processed by a non-linear machine learning algorithm. As a pre-processing step, we applied min-max scaling before linear regression. However, we did not apply any scaling for XGBoost since it is based on space-partitioning tree and hence is not sensitive to scale [92]. Regarding the individual performance of feature sets, it can be seen that the covariates with linear regression perform better than speech features with XGBoost showing that the covariates possess important information about the level of the motor-related symptoms.

One of the advantages of conducting LOSO cross-validation is that LOSO schema utilizes the maximum amount of data in the training process. We further examined the influence of training set size on the performance of the XGBoost algorithm. In particular, we compared the results obtained by applying LOSO cross-validation with two other cross-validation schema where the data is split into training and testing sets with ratios $\%75 - \%25$ and $\%50 - \%25$. This comparison was performed in the following way: we first randomly selected %75 of the subjects for training and the remaining %25 was used for testing. Then, the $\%50 - \%25$ split was obtained by selecting the $\sim \%66$ of the previously selected training set, and using the same test

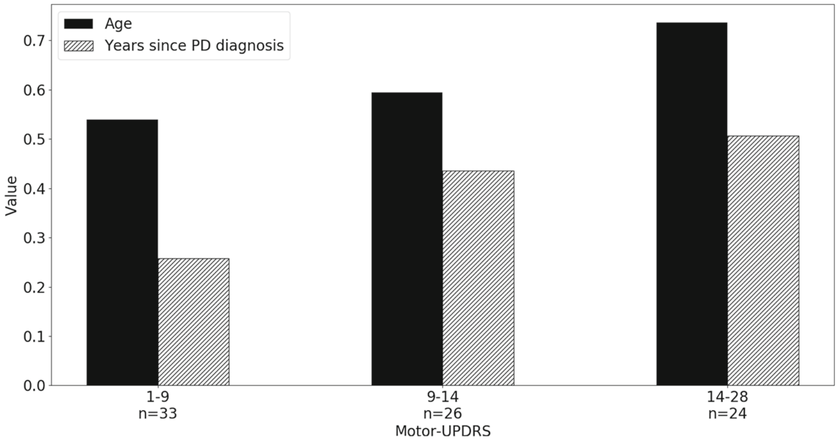


Fig. 1 Scaled values of age and years since PD diagnosis grouped by patients' motor-UPDRS scores

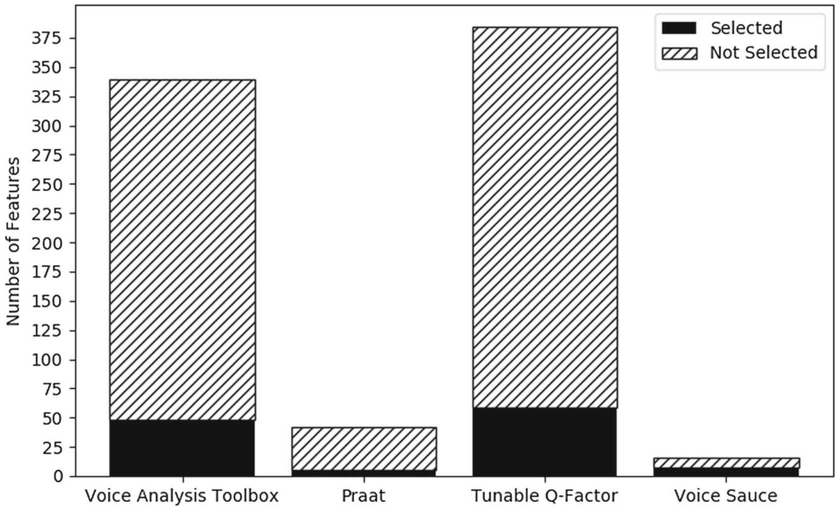


Fig. 2 Number of features of speech signal processing tools

set that was obtained in the previous step. This procedure ensures that the test set is always the same for both ratios. This process has been repeated 100 times with the setting where motor-UPDRS is predicted with XGBoost using speech data and covariates, and the averages of the obtained results are displayed in Table 7. The results indicate that the performance of the algorithm improves as more data is used in the training process.

4.3 Results on Parkinson's telemonitoring dataset

We applied the approach proposed in this study on the publicly available Parkinson's telemonitoring dataset [77] to compare its performance to the best result achieved on this dataset in the literature which used an unbiased cross-validation method. To the best of our knowledge, the study by Naranjo et al. [47] has reported the best results on Parkinson's telemonitoring dataset by using an unbiased cross-validation method. Naranjo et al. randomly split the recordings of approximately %75 percent of the subjects for training and the rest for testing and applied a Bayesian linear regression for performing the UPDRS estimation. They repeated this process 100 times and then reported the average of the obtained results. We followed a similar procedure with the difference that we performed a feature selection step with Boruta and the UPDRS predictions were performed with XGBoost algorithm. Table 8 shows the obtained MAEs and standard deviation with our approach on the Parkinson's telemonitoring dataset along with the results reported in [47]. We provide overall and also the

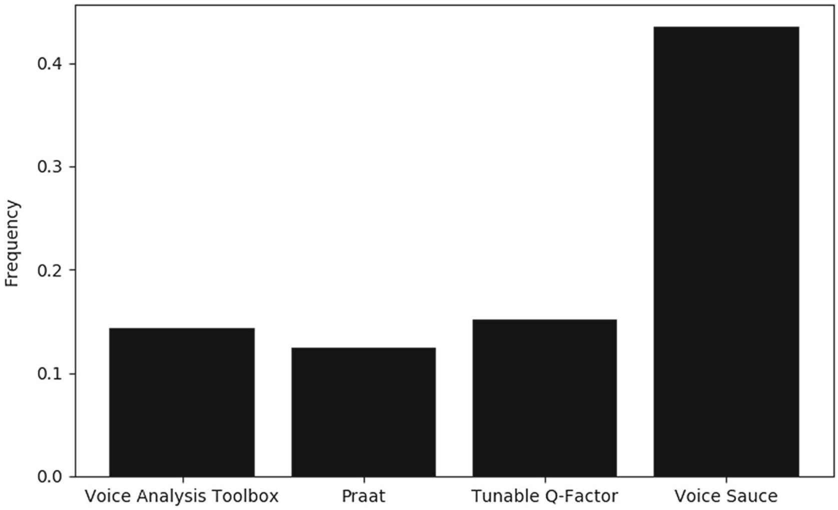


Fig. 3 Feature occurrence frequencies of speech signal processing tools

Table 6 Results of the motor-UPDRS (0-56 scale) predictions obtained with the proposed approach

Feature set	MAE	Spearman
Covariates with linear regression	4.01	0.38
Speech with XGBoost	4.14	0.35
(*) Speech + covariates (XGBoost)	3.96	0.42
(**) Speech (XGBoost) + covariates (linear regression)	3.87	0.46
(*) Data-level fusion, (**) prediction level fusion		

gender-specific results in Table 8. The results demonstrate that the proposed approach resulted in a lower MAE in all three cases.

5 Conclusions and discussion

In this study, we address the problem of disease severity assessment and symptom monitoring in Parkinson's disease (PD) investigations. We analyze the sustained vowel recordings of PD patients and propose a machine learning framework to predict the motor-UPDRS section. We present our findings on a recent dataset collected in the context of our project [65] and another public dataset.

Considering that telemonitoring applications are designed to be used for patient follow-up on a regular basis and hence patients' demographic information and case history are available, we have incorporated demographic and patient information such as age, gender, and years since diagnosis into the dysphonia-based estimation model. The results showed that using these covariates together with speech features improves the accuracy of the telemonitoring system. The additional covariates are used as a kind of correction factor by the estimation model since they are highly correlated with the progress of the symptoms. The prediction level fusion, in which two independent models are trained and then their predictions are combined, gave lower error than data-level fusion, in which covariates and speech features are first merged and then a single model is trained on the obtained feature set. The feature set analysis with the wrapper-based feature selection algorithm, called Boruta, showed that the features related to the vibration pattern of the vocal fold are effective features in UPDRS estimation.

Table 7 Comparison of leave-one-subject-out with %75 − %25 and %50 − %25 splits

Feature set	MAE	Spearman
Leave-one-subject-out	3.96	0.42
%75 − %25	4.24	0.38
%50 − %25	4.28	0.36

In the context of PD telemonitoring studies based on speech tests, the main issue this and many other studies have identified is the use of a non-subject-wise CV scheme. In many cases, subjects undergo a repeated measurement of the same test at the same time. If a subject-wise CV scheme is not used in the presence of this type of data, some of the measurements of a subject may remain in the training set while the rest occur in the test set. This could potentially introduce a dependence between training and test sets. It has been pointed out by many studies that the use of improper CV methods could produce overly misleading results [5, 47, 62–64, 83]. Besides, feature subset selection bias, which occurs when all the samples are used in the selection of a subset of features, should also be avoided to construct a realistic telemonitoring system. In this study, we aimed to apply a totally unbiased machine learning framework by using proper cross-validation methods during both the feature selection and model training steps. The best results obtained under the unbiased realistic setting in our experiments, mean absolute error of 3.87 and Spearman correlation of 0.46, showed the potential of the use of such telemonitoring applications for PD severity assessment.

There are some studies which reported that the sustained vowel test may not be an appropriate assessment tool for PD. In one of these studies, Bayestehtashk et al. [5] reported that the diadochokinetic task and particularly the reading task were better able to predict the motor-UPDRS

Table 8 Comparison of the motor-UPDRS (0–108 scale) prediction results on the Parkinson's telemonitoring dataset [77]

Study	Features	MAE
This study	Speech	7.13±1.07
	Speech (males)	7.89±1.50
	Speech (females)	6.91±1.62
Naranjo et al. [47]	Speech	7.52±1.10
	Speech (males)	8.22±1.49
	Speech (females)	8.79±2.89

section than the sustained phonation task. Our motivation for conducting the sustained vowel test was that it is a simpler test which can be feasible in real-life applications. Also, the analysis conducted by Bayestehtashk et al. [5] was based only on the linear learning models. It was previously discussed by Tsanas et al. [77, 78] that the linear learning models may have limitations and they suggested the use of nonlinear models for the prediction of UPDRS with features extracted from sustained phonations. In another related study, Lipsmeier et al. [36] conducted a study where the participants (44 with PD and 35 healthy controls) performed six tests (sustained phonation, rest tremor, postural tremor, finger tapping, balance, and gait) with their smartphones. They reported that except for the sustained phonation test, the features extracted from the other tests are significantly correlated with the corresponding assessments in the UPDRS. The reason for not observing a significant correlation between the sustained phonation test and the speech assessments in the UPDRS might be because they only extracted a single feature, namely a mel-frequency cepstral coefficient, from the data captured by the sustained phonation test. Also, even though UPDRS is considered to have good inter-rater variability overall [71], some elements in UPDRS, which includes speech assessments, display high inter-rater variability [22, 49, 60].

The literature studies revealed that the motor-UPDRS section of UPDRS has approximately four to five points [57] inter-rater variability in average (on 0–108 scale). Based on this observation, an error close to these values can be seen as an effective decision support system [47, 78]. However, in this study, we collected data from a single clinic, so as discussed by Bayestehtashk et al. [5], clinic-specific practices might have an influence on our results. For instance, Bayestehtashk et al. [5] used the data collected from three clinics and they included a clinic-wise analysis in their study. They used the data collected from one clinic for training and evaluated their results on the data obtained from the other two clinics. Besides, in our study, UPDRS assessments at each session were conducted by one of the neurologists in the project group. To circumvent these limitations, in future work, the data can be collected from multiple clinics and also multiple neurologist experts can apply UPDRS to each patient. We believe that using average UPDRS score from multiple neurologist experts might decrease the teacher noise and thus increase the reliability of the obtained model.

Funding This research was supported by The Scientific and Technological Research Council of Turkey (TUBITAK) grant number 215E008, https://www.tubitak.gov.tr/en.

Compliance with ethical standards The study has the approval of the Clinical Research Ethics Committee of Bahcesehir University.

References

1. Abdulhay E, Arunkumar N, Narasimhan K, Vellaiappan E, Venkatraman V (2018) Gait and tremor investigation using machine learning techniques for the diagnosis of Parkinson disease. Fut Gener Comput Syst 83:366–373
2. Afonso LC, Rosa GH, Pereira CR, Weber SA, Hook C, Albuquerque VHC, Papa JP (2019) A recurrence plot-based approach for Parkinson's disease identification. Fut Gener Comput Syst 94:282–292
3. Al Mamun KA, Alhussein M, Sailunaz K, Islam MS (2017) Cloud based framework for Parkinson's disease diagnosis and monitoring system for remote healthcare applications. Fut Gener Comput Syst 66:36–47
4. Asuncion A, Newman D (2007) UCI machine learning repository
5. Bayestehtashk A, Asgari M, Shafran I, McNames J (2015) Fully automated assessment of the severity of Parkinson's disease from speech. Comput Speech Lang 29(1):172–185
6. Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. J Mach Learn Res 13(Feb):281–305
7. Boersma P (2006) Praat: doing phonetics by computer. http://www.praat.org/
8. Buza K, Ágnes Varga N (2016) ParkinsoNET: Estimation of UPDRS score using hubness-aware feedforward neural networks. Appl Artif Intell 30(6):541–555
9. Chan P, Holford N (2001) Drug treatment effects on disease progression. Ann Rev Pharmacol Toxicol 41(1):625–659
10. Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, pp 785–794
11. Chen X, Huang L, Xie D, Zhao Q (2018) EGBMMDA: extreme gradient boosting machine for miRNA-disease association prediction. Cell Death Disease 9(1):3
12. Chiba T, Kajiyama M (1958) The vowel: its nature and structure, vol 652. Phonetic society of Japan Tokyo
13. Dorsey E, Constantinescu R, Thompson J, Biglan K, Holloway R, Kieburtz K, Marshall F, Ravina B, Schifitto G, Siderowf A et al (2007) Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. Neurology 68(5):384–386
14. Duffy J (2013) Motor speech disorders: substrates, differential diagnosis, and management, 3 edn. Elsevier Mosby, St Louis
15. Espay AJ, Bonato P, Nahab FB, Maetzler W, Dean JM, Klucken J, Eskofier BM, Merola A, Horak F, Lang AE, Reilmann R, Giuffrida J, Nieuwboer A, Horne M, Little MA, Litvan I, Simuni T, Dorsey ER, Burack MA, Kubota K, Kamondi A, Godinho C, Daneault J, Mitsi G, Krinke L, Hausdorff JM, Bloem BR, Papapetropoulos S (2016) Technology in Parkinson's disease: challenges and opportunities. Mov Disord 31(9):1272–1282
16. Fahn S, Elton R et al (1987) Unified parkinson's disease rating scale. Recent Dev Parkinson's Disease 2:153–164
17. Friedman J, Hastie T, Tibshirani R (2001) The elements of statistical learning, vol 1. Springer series in statistics, New York
18. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Annals of statistics:1189–1232
19. Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38(4):367–378
20. Gelzinis A, Verikas A, Bacauskiene M (2008) Automated speech analysis applied to laryngeal disease categorization. Comput Methods Prog Biomed 91(1):36–47
21. Godino-Llorente JI, Gomez-Vilda P, Blanco-Velasco M (2006) Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters. IEEE Trans Biomed Eng 53(10):1943–1953

22. Goetz CG, Stebbins GT (2004) Assuring interrater reliability for the UPDRS motor section: utility of the UPDRS teaching tape. Movement Disord 19(12):1453–1456
23. Goetz CG, Stebbins GT, Wolff D, DeLeeuw W, Bronte-Stewart H, Elble R, Hallett M, Nutt J, Ramig L, Sanger T et al (2009) Testing objective measures of motor impairment in early Parkinson's disease: feasibility study of an at-home testing device. Mov Disord 24(4):551–556
24. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
25. Harel B, Cannizzaro M, Snyder PJ (2004) Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: a longitudinal case study. Brain Cogn 56(1):24–29
26. Hartelius L, Svensson P (1994) Speech and swallowing symptoms associated with Parkinson's disease and multiple sclerosis: a survey. Folia Phoniatrica Logopaedica 46(1):9–17
27. Herd CP, Tomlinson CL, Deane KH, Brady MC, Smith CH, Sackley CM, Clarke CE (2012) Speech and language therapy versus placebo or no intervention for speech problems in Parkinson's disease. The Cochrane Library
28. Hindle JV (2010) Ageing, neurodegeneration and Parkinson's disease. Age Ageing 39(2):156–161
29. Ho AK, Iansek R, Marigliani C, Bradshaw JL, Gates S (1999) Speech impairment in a large sample of patients with Parkinson's disease. Behav Neurol 11(3):131–137
30. Jacobson BH, Johnson A, Grywalski C, Silbergleit A, Jacobson G, Benninger MS, Newman CW (1997) The voice handicap index (VHI): development and validation. Amer J Speech-Lang Pathol 6(3):66–70
31. Jankovic J (2008) Parkinson's disease: clinical features and diagnosis. Journal of Neurology. Neurosur Psych 79(4):368–376
32. Kowal SL, Dall TM, Chakrabarti R, Storm MV, Jain A (2013) The current and projected economic burden of Parkinson's disease in the United States. Mov Disord 28(3):311–318
33. Kumar M, Pachori RB, Acharya UR (2017) Characterization of coronary artery disease using flexible analytic wavelet transform applied on ECG signals. Biomed Signal Process Control 31:301–308
34. Kursa M, Jankowski A, Rudnicki W (2010) Boruta – a system for feature selection. Fund Inf 101:271–285
35. Kursa M, Rudnicki W (2010) Feature selection with the boruta package. J Stat Softw 36(11):1–13
36. Lipsmeier F, Taylor KI, Kilchenmann T, Wolf D, Scotland A, Schjodt-Eriksen J, Cheng WY, Fernandez-Garcia I, Siebourg-Polster J, Jin L et al (2018) Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. Mov Disord 33(8):1287–1297
37. Logemann JA, Fisher HB, Boshes B, Blonsky ER (1978) Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. J Speech Hear Disord 43(1):47–57
38. Magdalinou N, Morris HR (2017) Clinical features and differential diagnosis of parkinson's disease. In: Movement disorders curricula. Springer, pp 103–115
39. Majdinasab F, Karkheiran S, Moradi N, Shahidi GA, Salehi M (2012) Relation between Voice Handicap Index (VHI) and disease severity in Iranian patients with Parkinson's disease. Med J Islamic Republ Iran 26(4):157
40. Markaki M, Stylianou Y, Arias-Londoño JD, Godino-Llorente JI (2010) Dysphonia detection based on modulation spectral features and cepstral coefficients. In: 2010 IEEE international conference on Acoustics speech and signal processing (ICASSP). IEEE, pp 5162–5165
41. Mekyska J, Janousova E, Gomez-Vilda P, Smekal Z, Rektorova I, Eliasova I, Kostalova M, Mrackova M, Alonso-Hernandez JB, Faundez-Zanuy M et al (2015) Robust and complex approach of pathological speech signal analysis. Neurocomputing 167:94–111
42. Mekyska J, Rektorova I, Smekal Z (2011) Selection of optimal parameters for automatic analysis of speech disorders in Parkinson's disease. In: 2011 34th international conference on Telecommunications and signal processing (TSP). IEEE, pp 408–412
43. Michaelis D, Gramss T, Strube HW (1997) Glottal-to-noise excitation ratio–a new measure for describing pathological voices. Acta Acust United Acust 83(4):700–706
44. Midi I, Dogan M, Koseoglu M, Can G, Sehitoglu MA, Gunal DI (2008) Voice abnormalities and their relation with motor dysfunction in Parkinson's disease. Acta Neurol Scand 117(1):26–34
45. Miller A (2002) Subset selection in regression. CRC Press, Boca Raton
46. Murty KSR, Yegnanarayana B (2008) Epoch extraction from speech signals. IEEE Trans Audio Speech Lang Process 16(8):1602–1613
47. Naranjo L, Pérez CJ, Martín J (2017) Addressing voice recording replications for tracking Parkinson's disease progression. Med Biol Eng Comput 55(3):365–373
48. Ozkanca Y, Goksu Ozturk M, Ekmekci MN, Atkins DC, Demiroglu C, Hosseini Ghomi R (2019) Depression screening from voice samples of patients affected by parkinson's disease. Digit Biomarkers 3(2):72–82
49. Mart?nez-Martín P, Gil-Nagel A, Gracia LM, Gómez JB, Martínez-Sarriés J, Bermejo F (1994) Unified Parkinson's disease rating scale characteristics and structure. Mov Disord 9(1):76–83
50. Pahwa R, Lyons KE (2013) Handbook of Parkinson's disease. Crc Press, Boca Raton
51. Paja W, Wrzesień M (2013) Melanoma important features selection using random forest approach. In: 2013 6Th international conference on human system interactions (HSI). IEEE, pp 415–418
52. Patidar S, Pachori RB, Acharya UR (2015) Automated diagnosis of coronary artery disease using tunable-Q wavelet transform applied on heart rate signals. Knowl-Based Syst 82:1–10
53. Patidar S, Pachori RB, Upadhyay A, Acharya UR (2017) An integrated alcoholic index using tunable-Q wavelet transform based features extracted from EEG signals for diagnosis of alcoholism. Appl Soft Comput 50:71–78
54. Perry TL, Ohde RN, Ashmead DH (2001) The acoustic bases for gender identification from children's voices. J Acoust Soc Amer 109(6):2988–2998
55. Poona NK, Ismail R (2014) Using Boruta-selected spectroscopic wavebands for the asymptomatic detection of Fusarium circinatum stress. IEEE J Sel Top Appl Earth Observ Remote Sens 7(9):3764–3772
56. Poona NK, Van Niekerk A, Nadel RL, Ismail R (2016) Random forest (RF) wrappers for waveband selection and classification of hyperspectral data. Appl Spectroscopy 70(2):322–333
57. Post B, Merkus MP, de Bie R, de Haan RJ, Speelman JD (2005) Unified Parkinson's disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable? Movement Disord 20(12):1577–1584
58. Rahn DA, Chou M, Jiang JJ, Zhang Y (2007) Phonatory impairment in parkinson's disease: evidence from nonlinear dynamic analysis and perturbation analysis. J Voice 21(1):64–71
59. Ramaker C, Marinus J, Stiggelbout AM, Van Hilten BJ (2002) Systematic evaluation of rating scales for impairment and disability in Parkinson's disease. Mov Disord 17(5):867–876

60. Richards M, Marder K, Cote L, Mayeux R (1994) Interrater reliability of the Unified Parkinson's Disease Rating Scale motor examination. Mov Disord 9(1):89–91
61. Rusz J, Cmejla R, Ruzickova H, Ruzicka E (2011) Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. J Acoust Soc Amer 129(1):350–367
62. Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP (2017) The need to approximate the use-case in clinical machine learning. GigaScience 6(5):1–9
63. Sakar BE, Isenkul ME, Sakar CO, Sertbas A, Gurgen F, Delil S, Apaydin H, Kursun O (2013) Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. IEEE J Biomed Health Inf 17(4):828–834
64. Sakar CO, Kursun O (2010) Telediagnosis of Parkinson's disease using measurements of dysphonia. J Med Syst 34(4):591–599
65. Sakar CO, Serbes G, Gunduz A, Tunc HC, Nizam H, Sakar BE, Tutuncu M, Aydin T, Isenkul ME, Apaydin H (2019) 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 74:255–263
66. Schoentgen J, De Guchteneere R (1995) Time series analysis of jitter. J Phon 23(1-2):189–201
67. Schüpbach WMM, Corvol JC, Czernecki V, Djebara MB, Golmard JL, Agid Y, Hartmann A (2010) Segmental progression of early untreated Parkinson's disease: a novel approach to clinical rating. J Neurol Neurosurg Psych 81(1):20–25
68. Selesnick IW (2011) Wavelet transform with tunable Q-factor. IEEE Trans Signal Process 59(8):3560–3575
69. Serbes G, Sakar BE, Gulcur HO, Aydin N (2015) An emboli detection system based on Dual Tree Complex Wavelet Transform and ensemble learning. Appl Soft Comput 37:87–94
70. Shue YL (2010) The voice source in speech production: data, analysis and models. Ph.D. thesis, University of California, Los Angeles
71. Siderowf A, McDermott M, Kieburtz K, Blindauer K, Plumb S, Shoulson I, Group PS (2002) Test–retest reliability of the unified Parkinson's disease rating scale in patients with early Parkinson's disease: results from a multicenter clinical trial. Movement Disord 17(4):758–763
72. Simpson AP (2009) Phonetic differences between male and female speech. Lang Linguist Compass 3(2):621–640
73. Singhi SK, Liu H (2006) Feature subset selection bias for classification learning. In: Proceedings of the 23rd international conference on Machine learning. ACM, pp 849–856
74. Sánchez-Ferro Á, Elshehabi M, Godinho C, Salkovic D, Hobert MA, Domingos J, Uem JM, Ferreira JJ, Maetzler W (2016) New methods for the assessment of Parkinson's disease (2005 to 2015): A systematic review. Mov Disord 31(9):1283–1292
75. Touzani S, Granderson J, Fernandes S (2018) Gradient boosting machine for modeling the energy consumption of commercial buildings. Energy Build 158:1533–1543
76. Tsanas A (2012) Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning. Ph.D. thesis, Oxford Centre for Industrial and Applied Mathematics, University of Oxford, UK
77. Tsanas A, Little MA, McSharry PE, Ramig LO (2010) Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. IEEE Trans Biomed Eng 57(4):884–893
78. Tsanas A, Little MA, McSharry PE, Ramig LO (2011) Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. J R Soc Interface 8(59):842–855
79. Tsanas A, Little MA, McSharry PE, Spielman J, Ramig LO (2012) Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Trans Biomed Eng 59(5):1264–1271
80. Tsanas A, Ma L, McSharry P, Ramig L (2010) New nonlinear markers and insights into speech signal degradation for effective tracking of Parkinson's disease symptom severity. In: New nonlinear markers and insights into speech signal degradation for effective tracking of parkinson's disease symptom severity, Krakow, pp 457–460
81. Ulukaya S, Serbes G, Kahya YP (2017) Overcomplete discrete wavelet transform based respiratory sound discrimination with feature and decision level fusion. Biomed Signal Process Control 38:322–336
82. Ulukaya S, Serbes G, Kahya YP (2019) Wheeze type classification using non-dyadic wavelet transform based optimal energy ratio technique. Comput Biol Med 104:175–182
83. Vaiciukynas E, Verikas A, Gelzinis A, Bacauskiene M (2017) Detecting Parkinson's disease from sustained phonation and speech signals. Plos One 12(10):e0185613
84. Van Den Eeden SK, Tanner CM, Bernstein AL, Fross RD, Leimpeter A, Bloch DA, Nelson LM (2003) Incidence of Parkinson's disease: variation by age, gender, and race/ethnicity. Amer J Epidemiol 157(11):1015–1022
85. Vos T et al (2017) Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990—2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet 390(10100):1211–1259
86. Weismer G, Jeng JY, Laures JS, Kent RD, Kent JF (2001) Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders. Folia Phoniatrica Logopaedica 53(1):1–18
87. Wenning GK, Tison F, Seppi K, Sampaio C, Diem A, Yekhlef F, Ghorayeb I, Ory F, Galitzky M, Scaravilli T et al (2004) Development and validation of the unified multiple system atrophy rating scale (UMSARS). Mov Disord 19(12):1391–1402
88. Whitmore LS, Davis RW, McCormick RL, Gladden JM, Simmons BA, George A, Hudson CM (2016) BiocompoundML: a general biofuel property screening tool for biological molecules using Random Forest Classifiers. Energy Fuels 30(10):8410–8418
89. Zarzur AP, Duarte IS, Gonċalves GdNH, Martins MAUR (2010) Laryngeal electromyography and acoustic voice analysis in Parkinson's disease: a comparative study. Brazil J Otorhinolaryngol 76(1):40–43
90. Zhan A, Mohan S, Tarolli C et al (2018) Using smartphones and machine learning to quantify parkinson disease severity: The mobile parkinson disease score. JAMA Neurology
91. Zhang Y, Haghani A (2015) A gradient boosting method to improve travel time prediction. Transp Res Part C: Emerg Technol 58:308–324
92. Zheng A, Casari A (2018) Feature engineering for machine learning: principles and techniques for data scientists. O'Reilly Media, Inc.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

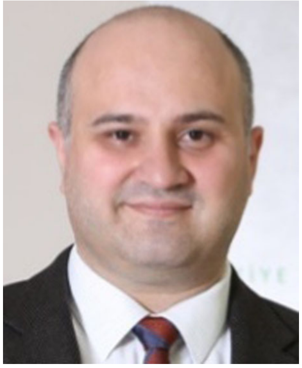
Hunkar Can Tunc is currently pursuing the M.S. degree at the University of Konstanz, Kostanz, Germany. His current research interests include machine learning and formal verification.

C. Okan Sakar has been working in the Computer Engineering Department of Bahcesehir University since 2014 as an assistant professor. He is interested in machine learning and pattern recognition.

Hulya Apaydin graduated from Istanbul University Cerrahpasa Medical Faculty Department of Neurology as Neurologist. She has been working in the same faculty since 1987. She was involved in several clinical researches.

Gorkem Serbes has experience in biomedical engineering, digital signal processing and pattern recognition for more than 10 years, and he has authored over 40 peer-reviewed manuscripts in these fields.

Aysegul Gunduz was graduated from Istanbul University, Cerrahpasa School of Medicine in 2003. She completed Neurology residency in the same faculty in 2008. She is interested in clinical neurophysiology and movement disorders.



Melih Tutuncu completed Neurology residency in Istanbul University, Cerrahpasa Faculty of Medicine in 2010. He has worked in Mayo Clinic, United States as a research fellow between 2010 and 2011. He is interested in clinical neurology.

Fikret Gurgen received the Ph.D. in electrical engineering from the University of Akron in 1989 and is currently a professor at Bogazici University, Turkey. He is interested in intelligent systems, medical decision-making applications, and speech processing.

Affiliations

Hunkar C. Tunc[1,2] · **C. Okan Sakar**[1] · **Hulya Apaydin**[3] · **Gorkem Serbes**[4] · **Aysegul Gunduz**[3] · **Melih Tutuncu**[3] · **Fikret Gurgen**[5]

C. Okan Sakar
okan.sakar@eng.bau.edu.tr

Hulya Apaydin
hulyapay@istanbul.edu.tr

Gorkem Serbes
gserbes@yildiz.edu.tr

Aysegul Gunduz
aysegul.gunduz@istanbul.edu.tr

Melih Tutuncu
tutuncumelih@yahoo.com

Fikret Gurgen
gurgen@boun.edu.tr

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

[2] Present address: Department of Computer and Information Science, University of Konstanz, Konstanz, Germany

[3] Department of Neurology, Cerrahpasa Medical Faculty, Istanbul University-Cerrahpasa, 34098, Istanbul, Turkey

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

[5] Department of Computer Engineering, Bogazici University, 34342, Istanbul, Turkey