Kodali, Manila; Kadiri, Sudarsana; Alku, Paavo

Automatic classification of the severity level of Parkinson’s disease: A comparison of speaking tasks, features, and classifiers

Published in:
Computer Speech and Language

DOI:
10.1016/j.csl.2023.101548

Published: 01/10/2023

Document Version
Publisher's PDF, also known as Version of record

Published under the following license:
CC BY

Please cite the original version:
Kodali, M., Kadiri, S., & Alku, P. (2023). Automatic classification of the severity level of Parkinson’s disease: A comparison of speaking tasks, features, and classifiers. *Computer Speech and Language*, *83*, Article 101548.
https://doi.org/10.1016/j.csl.2023.101548

Contents lists available at ScienceDirect

Computer Speech & Language

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

Automatic classification of the severity level of Parkinson's disease: A comparison of speaking tasks, features, and classifiers

Manila Kodali [*], Sudarsana Reddy Kadiri, Paavo Alku

Department of Information and Communications Engineering, Aalto University, Finland

ARTICLE INFO

Keywords:
Parkinson's disease
Severity level classification
MDS-UPDRS-III
MDS-UPDRS-S

ABSTRACT

Automatic speech-based severity level classification of Parkinson's disease (PD) enables objective assessment and earlier diagnosis. While many studies have been conducted on the binary classification task to distinguish speakers in PD from healthy controls (HCs), clearly fewer studies have addressed multi-class PD severity level classification problems. Furthermore, in studying the three main issues of speech-based classification systems – speaking tasks, features, and classifiers – previous investigations on the severity level classification have yielded inconclusive results due to the use of only a few, and sometimes just one, type of speaking task, feature, or classifier in each study. Hence, a systematic comparison is conducted in this study between different speaking tasks, features, and classifiers. Five speaking tasks (vowel task, sentence task, diadochokinetic (DDK) task, read text task, and monologue task), four features (phonation, articulation, prosody, and their fusion), and four classifier architectures (support vector machine (SVM), random forest (RF), multilayer perceptron (MLP), and AdaBoost) were compared. The classification task studied was a 3-class problem to classify PD severity level as healthy *vs.* mild *vs.* severe. Two MDS-UPDRS scales (MDS-UPDRS-III and MDS-UPDRS-S) were used for the ground truth severity level labels. The results showed that the use of the monologue task and the articulation and fusion of features improved classification accuracy significantly compared to the use of the other speaking tasks and features. The best classification systems resulted in a rate of accuracy of 58% (using the monologue task with the articulation features) for the MDS-UPDR-III scale and 56% (using the monologue task with fusion of features) for the MDS-UPDRS-S scale.

1. Introduction

Parkinson's disease (PD) is a neurological disorder affected by a progressive loss of nerve cells in the part of the brain called substantia nigra. PD is the second most prominent neurodegenerative disease next to Alzheimer's disease, which has a prevalence of about 0.5%–1% in people aged 65–69 years, increasing to 1%–3% in people aged 80 and over (Tanner and Goldman, 1996; Nussbaum and Ellis, 2003). People with PD suffer from many motor and non-motor impairments (Hornykiewicz, 1998). Motor impairments comprise bradykinesia, tremor, walking and gait difficulties, and speech disorders. Since there is no single test available to diagnose PD, medical experts detect PD and evaluate its severity level using a combination of various diagnostic tests.

In diagnosis of PD, the unified Parkinson's disease rating scale (UPDRS) is an extensively utilized clinical rating scale to evaluate the disease's progression (Jankovic, 2008; Nilashi et al., 2019; Arias-Vergara et al., 2018b; Wang et al., 2022). The initial version of UPDRS consists of 42 items, each item ranging from 0 (normal) to 4 (severe), adding up to a total scale of 199 points. The

[*] Corresponding author.
E-mail addresses: manila.kodali@aalto.fi (M. Kodali), sudarsana.kadiri@aalto.fi (S.R. Kadiri), paavo.alku@aalto.fi (P. Alku).

https://doi.org/10.1016/j.csl.2023.101548
Received 22 December 2022; Received in revised form 26 June 2023; Accepted 8 July 2023
Available online 16 July 2023

items are divided into four parts based on the characteristics they measure: Part I (mentation, behaviour, and mood; 4 items; 0–16 points), Part II (activities of daily living; 13 items; 0–52 points), Part III (motor examination; 14 items; 0–108 points), and Part IV (complications of therapy; 11 items; 0–23 points). The movement disorder society (MDS) proposed a newer version of the scale called the MDS-UPDRS. The MDS-UPDRS scale is more comprehensive, homogeneous, and clinimetrically superior to the UPDRS scale. The MDS-UPDRS comprises 65 items, each item ranging from 0 (normal) to 4 (severe), adding up to a total scale of 260 points. The items are divided into four parts, each of which differs slightly from its predecessor: Part I (non-motor aspects of experiences of daily living; 13 items; 0–52 points), Part II (motor activities of daily living; 13 items; 0–52 points), Part III (motor examination; 33 items; 0–132 points), and Part IV (motor complications; 6 items; 0–24 points) (Goetz et al., 2008). Since Part III addresses motor capabilities in PD, it is extensively used to rate the severity level of PD and is referred to as the MDS-UPDRS-III scale. The MDS-UPDRS-III has a full scale of 33 items, one of which is the speech item. The speech item is referred to as the MDS-UPDRS Speech (MDS-UPDRS-S) scale. The MDS-UPDRS-S scale evaluates volume, modulation (prosody) and clarity, including slurring, palilalia (repetition of syllables), and tachyphemia (rapid speech, running syllables together) (Goetz et al., 2008).

Diagnosis of PD using the different MDS-UPDRS scales described above calls for conducting various tests at a hospital by clinical experts. In these tests, the patient has to be with the examiner, which poses a limitation on the availability of diagnosis. Moreover, as the PD diagnosis procedures described above require evaluating several items by the clinical expert, diagnosis is subjective and therefore prone to intrinsic biases of experts due to factors such as the evaluator's experience and cognitive status. As a result of these issues, research has been conducted to improve the accuracy of PD diagnosis, to make the assessment more objective, to shorten the time it takes for the patient to reach a diagnosis, and avoid frequent visits to the clinics by the patient. In particular, several studies have been conducted to investigate novel objective biomarkers that could complement and augment present clinical procedures with technical tools based on recent advances in signal processing, and machine learning (ML). Non-invasive biomarkers, such as those obtained from medical imaging (Shinde et al., 2019), audio-visual analysis of motor cues (Oktay and Kocer, 2020) and speech (Moro-Velazquez et al., 2019a; Narendra et al., 2019), have been used increasingly to study PD. Speech-based, automatic biomarking of PD has particularly gained increasing interest because speech disorders are considered an important early sign of PD (Rusz et al., 2013; Hlavnička et al., 2017).

The speech-based classification of PD aims at automatically distinguishing speakers with PD from healthy controls (HCs) using the computer based on the acoustic speech signal. This binary classification problem has become a popular research topic in recent years, and many articles have been published on the topic (Orozco-Arroyave et al., 2016; Vásquez-Correa et al., 2017; Pérez-Toro et al., 2019; Erdogdu Sakar et al., 2017; Parisi et al., 2018; Moro-Velazquez et al., 2019b; Karan et al., 2021; Rusz et al., 2021a). (For more information about speech-based automatic classification of different disorders, the reader is referred to the reviews published in Rusz et al. (2021a), Hegde et al. (2019), Gómez-Vilda et al. (2022), Verde et al. (2018).) However, compared to the binary classification of PD, much less research has been conducted in the multi-classification of the PD severity level based on the MDS-UPDRS-III and MDS-UPDRS-S scales (Bocklet et al., 2013; Arias-Londoño and Gómez-García, 2019; Arias-Vergara et al., 2018a; Vásquez-Correa et al., 2018a). By utilizing the MDS-UPDRS-III and MDS-UPDRS-S scales, PD can be categorized into different severity levels (for example, mild *vs*. moderate *vs*. severe). Studying the classification of the severity level of PD, particularly between healthy voice and the mild level of the disease, helps to develop automatic speech-based technology for early detection of PD. In Bocklet et al. (2013), the classification of PD patients from HCs was studied, and the PD severity level (mild *vs*. moderate *vs*. severe) was estimated based on the UPDRS-III scale using a corpus of 88 patients and 88 controls. The authors extracted acoustic, prosodic, and glottal features from speech signals produced in different types of speaking tasks (syllable repetition tasks, read sentences, and paragraphs, as well as monologues). By using a support vector machine (SVM) classifier with PD severity level labels based on the UPDRS-III, the authors reported a balanced accuracy of 59.1%. In Arias-Londoño and Gómez-García (2019), the severity level of PD was studied by grouping patients and HCs into four categories based on the MDS-UPDRS-III scale: HCs (0 points), mildly affected patients (≤32 points), moderately affected patients (≤58 points), and severely affected patients (>59 points). Due to the small number of speech samples from patients with moderate and severe symptoms, these two classes were further combined into a single class, called affected patients. The experiments in Arias-Londoño and Gómez-García (2019) used the MDS-UPDRS-III scale from two corpora (Neurovoz (Moro-Velazquez et al., 2019a,b) and PC-GITA (Orozco-Arroyave et al., 2014)) to study an automatic 3-class PD severity level prediction problem (healthy *vs*. mild *vs*. affected). Three vowels (/a:/, /i:/, and /u:/) parameterized by perturbation, spectral, cepstral, and complexity features were used as speech data. Deep neural network (DNN) and convolutional neural network (CNN) architectures were used as classifiers by studying the effects of transfer learning and multimodal strategies. The best classification accuracy of 52% was obtained by using a multimodal deep learning approach. In Arias-Vergara et al. (2018a), automatic multi-class assessment was studied based on the MDS-UPDRS-S scale using a multi-class SVM following a one *vs*. all strategy. An extended version of the PC-GITA database (Orozco-Arroyave et al., 2014) was used as training data, and testing data were collected from 17 PD patients (9 male, 8 female) using the Apkinson mobile application. For the multi-class assessment, the authors used only prosody features extracted from the monologue task, which resulted in accuracy of 67% for training data, and 51% for testing data. In Vásquez-Correa et al. (2018a), the authors classified the patients into various stages of the disease based on the scales assigned by the neurologist. For MDS-UPDRS-III scale, the ranges per class were stated as follows: from 0 to 25 points (initial), from 25 to 50 points (intermediate), and higher than 50 points (severe). For MDS-UPDRS-S scale, scale 0 corresponded to the initial stage, 1 to the intermediate stage, and 2 or higher to the severe stage. From the various sounds produced during the continuous speech, onset and offset transitions were detected. These transitions were modelled using CNNs. The best classification result was obtained using speech onsets for both the MDS-UPDRS-III scale (accuracy of 37.8%) and the MDS-UPDRS-S scale (accuracy of 54.9%).

As described above, there are only a few previous studies (i.e., Bocklet et al. 2013, Arias-Londoño and Gómez-García 2019, Arias-Vergara et al. 2018a, Vásquez-Correa et al. 2018a) that have investigated the automatic severity level classification of PD

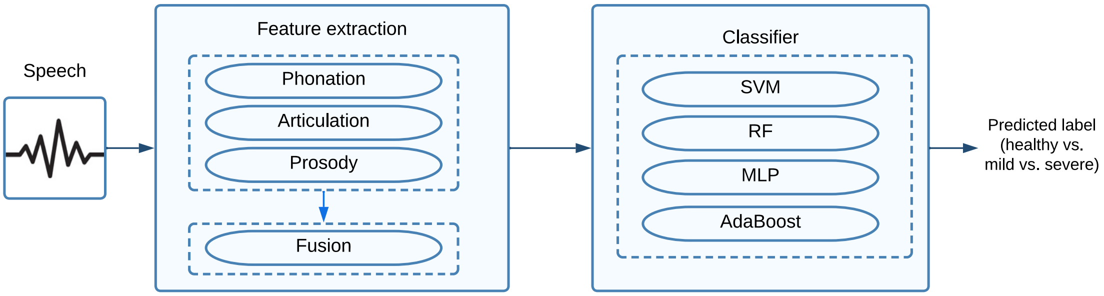


Fig. 1. A schematic block diagram of the PD severity level classification system used in this study.

from speech based on either MDS-UPDRS-III ratings or MDS-UPDRS-S ratings. However, these previous studies on the classification of the PD severity level have not specifically investigated the effect of speaking tasks, but each study has instead used individual speaking tasks such as the vowel task in Arias-Londoño and Gómez-García (2019), the monologue task in Arias-Vergara et al. (2018a), and the read text task in Vásquez-Correa et al. (2018a). Moreover, in some of the previous studies on the classification of the PD severity level (Arias-Vergara et al., 2018a), the topic has been studied only via the use of prosody features. In addition, all the previous studies listed above have used either the conventional SVM algorithm or the deep learning-based DNN, and CNN networks as classifier architectures without, however, comparing these architectures to each other.

Since prior research on speaking tasks, features, and classifiers in the automatic speech-based classification of the severity level of PD is scattered between a few individual investigations and is inconclusive, the present study aims to investigate these three issues systematically. More specifically, we will investigate five different speaking tasks (vowel task, sentence task, diadochokinetic (DDK) task, read text task, and monologue task) employing four types of features (phonation, articulation, prosody, and their fusion) by comparing four ML architectures (SVM, random forest (RF), multilayer perceptron (MLP), and AdaBoost). In this comparison, two standardized MDS-UPDRS scales (MDS-UPDRS-III and MDS-UPDRS-S) are used to score the neurological state of the patients. A total of three classes, *healthy* controls, PD patients with *mild* symptoms, and PD patients with *severe* symptoms, are studied according to the scales assigned by neurologists.

In short, the major contributions of this study are as follows :

- The automatic speech-based severity level classification of PD is investigated by systematically comparing different speaking tasks, features, and classifiers.
- In the automatic severity level classification, two labelling scenarios based on the MDS-UPDRS-III and MDS-UPDRS-S scales are compared.

The paper is organized as follows. Section 2 describes the methodology used including the feature extraction and the classifier architectures. The experiments are described in Section 3 including the database and evaluation metrics. Results are reported in Section 4, and the discussion and conclusions of the study are delineated in Section 5.

2. Methodology

The block diagram of the system used in this study to classify the PD severity level from speech is shown in Fig. 1. The system consists of two main stages: feature extraction and classification. In the feature extraction stage, four features (phonation, articulation, prosody, and the fusion of all three) are considered. In the classification stage, the input speech is classified as either *healthy* or *mild* or *severe* using an ML-based classifier. Four different ML classifiers (SVM, RF, MLP, and AdaBoost) that have been widely used in the field of speech-based biomarking of health (Hegde et al., 2019; Kadiri et al., 2020; Narendra and Alku, 2021) are compared in the study. These classifiers were selected because they are less data-hungry than deep learning-based models and, therefore, should work well for the small dataset used in this study. In addition, the selected classifiers are robust regarding outliers and can model complex interdependencies.

2.1. Feature extraction

Four features based on phonation, articulation, prosody, and their fusion are computed in the feature extraction. Extracting features that characterize phonation, articulation, and prosody is justified by previous studies showing that impairment of speech in PD is reflected by all these features, which helps in the automatic classification of PD (Garcia et al., 2017; López et al., 2019; Cernak et al., 2017). To capture vocal fold vibration's temporal and amplitudinal variations, phonation features are derived from voiced segments. Articulation features are derived using spectral measurements and speech energy at onset/offset transitions. Finally, the fundamental frequency (F0) contours and energy are used to derive prosody features. All of the features are extracted using the DisVoice framework (Orozco-Arroyave et al., 2018; Arias-Vergara et al., 2017; Vásquez-Correa et al., 2018b; Dehak et al., 2007).

2.1.1. Phonation

Phonation in speech production of PD patients is distinguished by bowing and inadequate closure of the vocal folds deteriorating stability and periodicity of vocal fold vibration. Therefore, for continuous speech, phonation features are derived from voiced segments. The phonation features consist of the following seven descriptors: the jitter, the shimmer, the first and second derivatives of F0, the amplitude perturbation quotient, the pitch perturbation quotient, and the energy. In addition, four statistical functionals (mean, standard deviation, skewness, and kurtosis) are computed per descriptor, resulting in one 28-dimensional phonation feature vector per utterance.

2.1.2. Articulation

Articulation deficits in PD patients are primarily due to the reduced amplitude and velocity in the lips, tongue, and jaw movements. The articulation features derived from the onset and offset of voiced segments consist of 12 Mel-frequency cepstral coefficients (MFCCs) along with their first and second derivatives and the signal's log energy computed in 22 Bark bands. In addition, the first and second formant frequencies, along with their derivatives, are used as articulation features. The total number of descriptors is 122. Next, four statistical functionals (mean, standard deviation, skewness, and kurtosis) are calculated for each descriptor, yielding one 488-dimensional articulation feature vector per utterance.

2.1.3. Prosody

Monotonicity, monoloudness, and variations in pauses and speech rate are impairments in the prosody of speech in PD (Rusz et al., 2011; Falk et al., 2012). The prosody features used in this study are based on duration as well as on the F0 and energy contour. The features related to duration are the voiced speech rate, the mean, the standard deviation of the duration of voiced segments, the pause rate, and the duration of pauses. The features related to the F0 are the mean, the standard deviation, and the maximum value of F0 and the variability of F0 represented in semitones. Finally, the features related to the energy are the mean, the standard deviation, and the maximum value of the energy contour. All these features yield one 103-dimensional prosody feature vector per utterance.

2.1.4. Fusion

In order to analyze whether there exists complementary information among the three features (phonation, articulation, and prosody) described in the previous sub-sections, all three features are concatenated, yielding one 619-dimensional feature vector per utterance, which is referred to as the fusion feature. Figs. 2 and 3 show examples of spectrograms, Mel-spectrograms and MFCCs computed for a sustained utterance of the vowel [a:] produced by a healthy female speaker and by two female PD patients who were classified to have mild and severe symptoms. In the examples shown in Figs. 2 and 3, the severity classification was based on the MDS-UPDRS-III and MDS-UPDRS-S respectively. It can be observed from the figures that there are differences between the three severity classes especially at low frequencies of the spectrograms and Mel-spectrograms and in the lowest MFCCs.

2.2. Classifier architectures

In this study, four different ML-based methods (SVM, RF, MLP, and AdaBoost) are used as the classifiers. In the following four sub-sections, the principles of these classifiers are briefly described.

2.2.1. Support vector machine (SVM)

SVM is a supervised ML classifier that transforms non-linear data into a high-dimensional space in which the data is linearly separable. SVM uses a hyper-plane to classify the classes distinctly. To determine the optimal hyper-plane, margins are maximized between the classes. Because of its great generalize capacity, SVM has been employed successfully in various classification problems (e.g., Kadiri et al. 2020, Shahbakhi et al. 2014). For more details about SVM, the reader is referred to Cortes and Vapnik (1995).

2.2.2. Random forest (RF)

RF is an ensemble-supervised ML classifier that constructs numerous decision trees and combines them to improve the accuracy of the classification system. The decision's complexity is reduced by checking only a subset of the features rather than all of them. To identify the best solution, the classifier generates many random decision trees. Each tree votes as a unit and the class with the most votes is chosen (Pal, 2005). RF is nonlinear, resistant to outliers, and prevents overfitting in high-dimensional data. For more details about RF, the reader is referred to Pal (2005).

2.2.3. Multi layer perceptron (MLP)

MLP is an artificial neural network (ANN) with multiple layers of neurons. In MLP, inputs and weights are multiplied for each neuron at the input layer. The net input is then transferred to the next layer using an activation function. The MLP is typically trained using the back propagation algorithm. This method's optimal weight values are computed iteratively using gradient descent. The change between the network output and the target output is defined as the error. The iteration of the MLP parameters is completed up to a given number of epochs, or until the reduction of the error has stopped (Pal and Mitra, 1992). Then, the network, whose weights and biases have been updated in the training phase, is used to predict the testing data. For more details about MLP, the reader is referred to Pal and Mitra (1992).

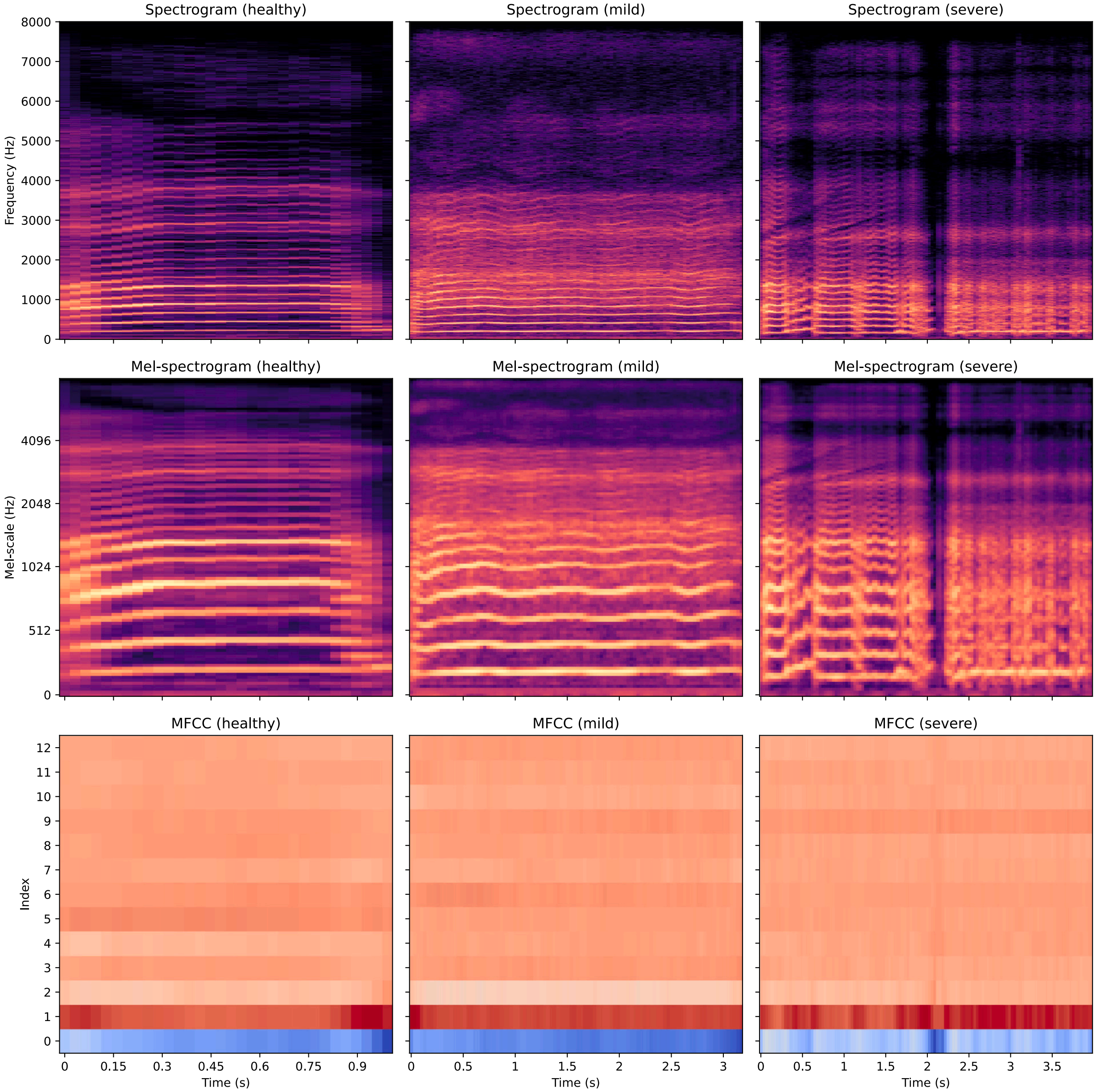


Fig. 2. Spectrograms, Mel-spectrograms and MFCCs computed for a sustained utterance of the vowel [a:] produced by a healthy female speaker (left column), by a female PD patient with mild symptoms (MDS-UPDRS-III = 19) (center column), and by a female PD patient with severe symptoms (MDS-UPDRS-III = 57) (right column).

2.2.4. AdaBoost

AdaBoost is an abbreviation for 'Adaptive Boosting'. AdaBoost is an effective boosting algorithm that aims to combine multiple weak learners to build one strong classifier. A weak hypothesis is produced by a weak learner. Each hypothesis is assigned a weight based on its accuracy. These weak hypotheses are combined to generate a strong hypothesis that is much more accurate than any of the rules (Schapire, 2013). For more details about AdaBoost, the reader is referred to Schapire (2013).

2.3. Optimization of the classifiers' hyper-parameters

The hyper-parameters of the classifiers are optimized using a grid search strategy where a set of possible parameter combinations are first formed for each of the classifiers and the optimal combination is then searched. The parameters of the grid search are

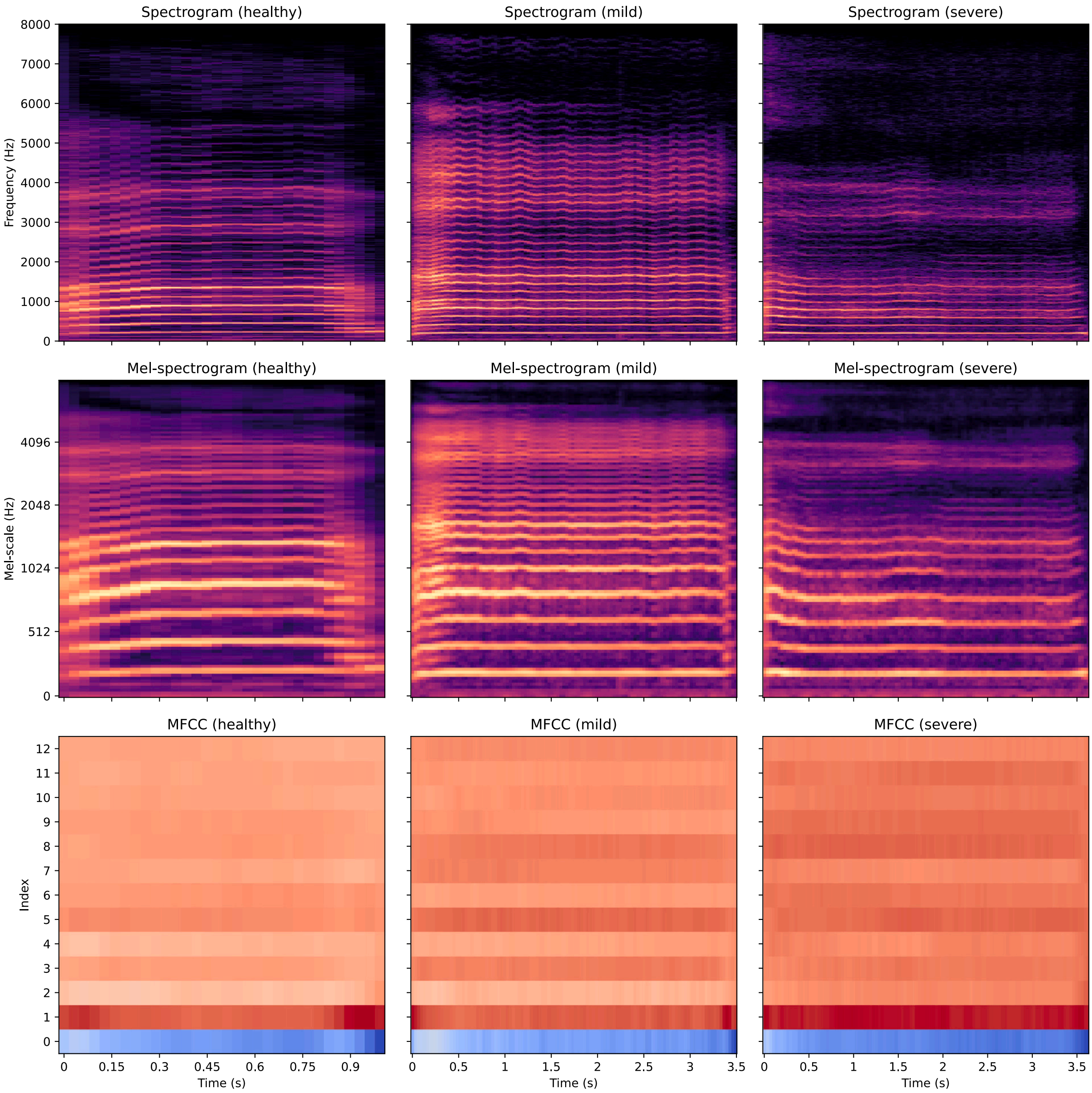


Fig. 3. Spectrograms, Mel-spectrograms and MFCCs computed for a sustained utterance of the vowel [a:] produced by a healthy female speaker (left column), by a female PD patient with mild symptoms (MDS-UPDRS-S = 0) (center column), and by a female PD patient with severe symptoms (MDS-UPDRS-S = 3) (right column).

presented in Table 1. In order to tune the hyperparameters, the nested cross-validation (CV) strategy was used. This CV method helps to tackle the over-fitting problem and the biased evaluation of the model selection by nesting the hyperparameter optimization procedure under the model selection procedure. Hence, this approach is also called double cross-validation. The inner loop (which is used for optimizing the hyper-parameters using the grid search) and the outer loop (which is used for estimating the performance of the models trained on the inner loops) were split based on the GroupKFold strategy with 5 folds. Due to the large number of the fitted hyper-parameters per each inner loop and each setup, the resulting optimal parameter values are not reported in the manuscript.

Table 1
Hyper-parameters used for the grid search.

Classifier	Hyper-parameters

SVM	c : {0.1, 1, 10, 100}
	gamma:{0.001, 0.01, 0.1, 1}
	kernel:{'rbf', 'poly', 'sigmoid'}

RF	estimators: {50, 150, 250}
	max_features: {'sqrt', 0.25, 0.5, 0.75, 1.0}
	min_samples_split: {2, 4, 6}

MLP	hidden_layers: {(50, 50, 50), (100,), (50, 100, 50)}
	activation: {'tanh', 'relu'}
	solver: {'sgd', 'adam', 'lbfgs'}
	alpha: {0.0001, 0.05}
	learning_rate: {'constant', 'adaptive'}

AdaBoost	estimators: {50, 150, 250}
	learning_rate: {0.001, 0.01, 0.1}

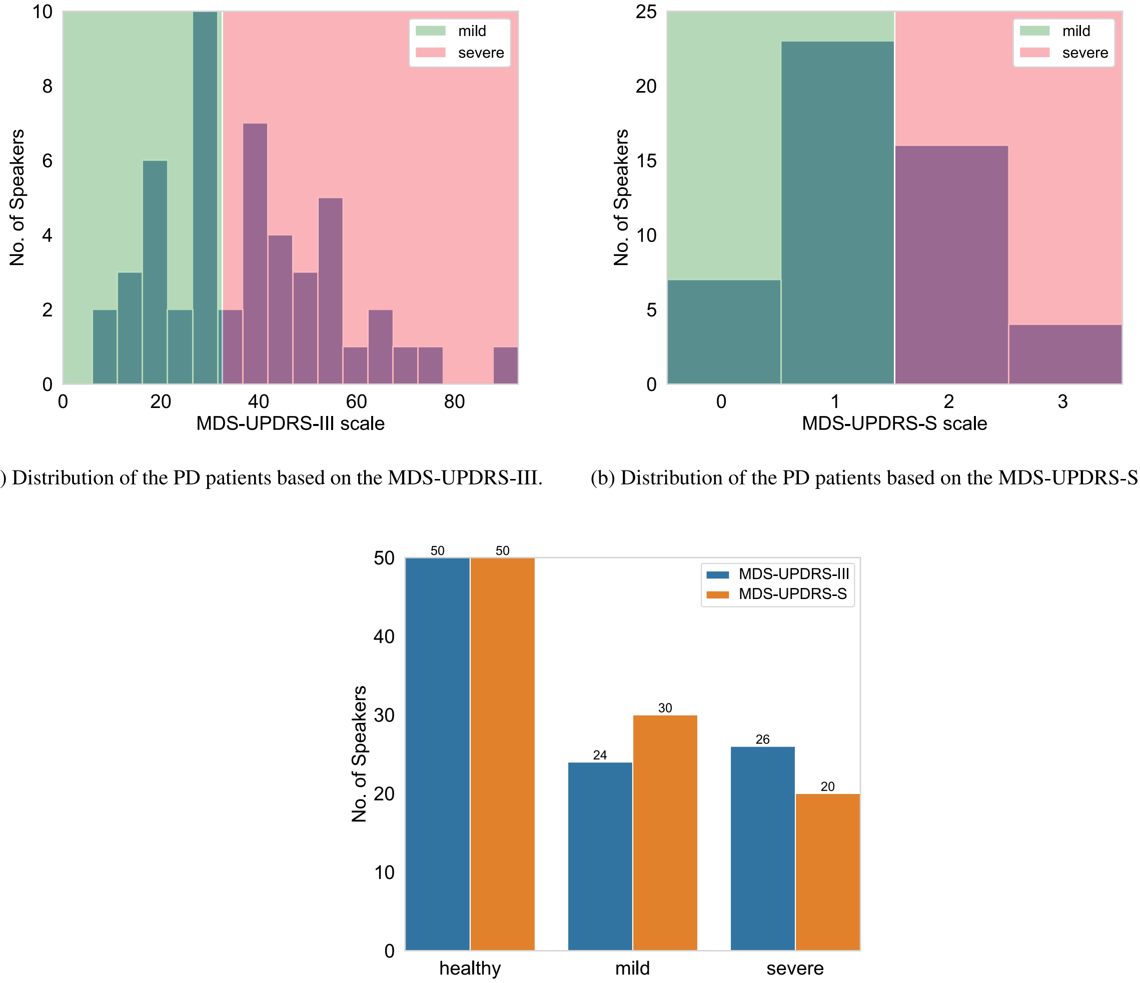
3. Experiments

3.1. Database

The data studied in this investigation is taken from the Spanish PD database PC-GITA (Orozco-Arroyave et al., 2014). This database is well suited for the present study because the data has been labelled using both the MDS-UPDRS-III and the MDS-UPDRS-S scale by neurologist experts. PC-GITA contains speech recordings from 50 PD patients and 50 HCs. The database is balanced with regards to gender and age. The disease duration after diagnosis for individuals with PD ranges from 0 to 43 years, with an average duration of 11 ± 10 years. The speech samples were recorded when the patients were in the ON-state, i.e., no more than 3 h after their morning medication. The database includes speech recordings collected using different speaking tasks. From the speaking tasks of PC-GITA, the following five tasks are studied in the current investigation: (1) production of sustained vowels, which will be referred to as the vowel task, (2) production of isolated sentences, which will be referred to as the sentence task, (3) rapid repetition of words (/pa-ta-ka/, /pa-ka-ta/, /pe-ta-ka/) and syllables (/pa/, /ta/, /ka/), which will be referred to as the DDK task, (4) reading of a dialog between a doctor and a patient, which will be referred to as the read text task, and (5) talking about different topics such as hobbies, daily living activities, family, and others, which will be referred to as the monologue task. In the vowel task, five vowels (/a/, /e/, /i/, /o/, and /u/) were included and repeated three times per speaker, resulting in 750 vowel sounds produced by the PD patients and 750 vowel sounds produced by the HCs. The sentence task consists of six different sentences resulting in 300 sentences produced by the PD patients and 300 sentences produced by the HCs. The DDK task includes 300 and 300 DDK utterances generated by PD patients and HCs, respectively. In the read text task, 50 PD patients and 50 HCs read aloud the same text describing a dialog between a doctor and a patient. The speech generated in the read text task was phonetically balanced. Finally, in the monologue task the speakers were asked to tell what they typically do on a normal day. This task consists of 50 monologues spoken by the PD patients and 50 monologues spoken by the HCs. The average duration for the vowel task was 3 ± 2 s, sentence task was 3 ± 1 s, DDK task was 4 ± 2 s, read text task was 18 ± 5 s, and monologue was 47 ± 26 s. The sampling frequency of the original recordings was 44.1 kHz, but the data was down-sampled to 16 kHz in this study. More details about the dataset can be found in Orozco-Arroyave et al. (2014).

The scale distribution for the complete MDS-UPDRS-III and MDS-UPDRS-S are shown in Figs. 4(a) and 4(b). The total MDS-UPDRS-III scale in the PG-GITA database ranges from 6 to 93 points. The MDS-UPDRS-S scale generally ranges from 0 to 4, which resembles to just one item among the 33 items of the MDS-UPDRS-III (Goetz et al., 2008). However, the MDS-UPDRS-S scale in the PG-GITA database ranges from 0 to 3. Two classes are determined from each histogram to conduct multi-class experiments to classify mild and severe stages of the disease. The ranges per class for the MDS-UPDRS-III scale are defined as follows: from 0 to 32 (mild), and higher than 32 (severe). For the MDS-UPDRS-S scale, we considered 0 and 1 to be mild, and 2 or higher as severe. The distribution and limits of the scale for each class are illustrated in Figs. 4(a) and 4(b). Due to the limited amount of the data, the division of the classes was grouped as described above. Similar categorizations have been used previously, for example, in Arias-Londoño and Gómez-García (2019). Note that a patient can in principle be in different classes depending on whether his or her disease severity level is rated according to the MDS-UPDRS-III or MDS-UPDRS-S scales (e.g., the same patient could be in the mild stage based on the MDS-UPDRS-III scale and in the severe stage based on the MDS-UPDRS-S scale). In order to further demonstrate how the PD patients of the current study are divided according to the MDS-UPDRS-III and MDS-UPDRS-S severity level labels, Fig. 5 shows the confusion matrix for the mild and severe classes labelled using the two MDS-UPDRS scales. The MDS-UPDRS scores were not collected for the HCs. Therefore, the utterances of the HCs were simply labelled as "healthy".

Taken together, the speakers of the PC-GITA database were divided for the purposes of the present study into three severity level classes (*healthy*, *mild*, and *severe*) based on their MDS-UPDRS-III and MDS-UPDRS-S scales. The number of speakers in each class is shown in Fig. 4(c). Gender and age distributions were balanced in the ML experiments. Further information about the impact of age and gender on speech in PD can be found in recent phenotypic studies (Rusz et al., 2021b, 2022).



(a) Distribution of the PD patients based on the MDS-UPDRS-III.

(b) Distribution of the PD patients based on the MDS-UPDRS-S.

(c) Number of speakers in the three PD severity level classes based on the MDS-UPDRS-III and MDS-UPDRS-S.

Fig. 4. (a) Distribution of the MDS-UPDRS-III values among PD patients. (b) Distribution of the MDS-UPDRS-S values among PD patients. In (a) and (b), patients in the mild stage are shown in green bars, and patients in the severe stage are shown in red bars. (c) Distribution of the classes healthy, mild, and severe among the MDS-UPDRS-III and MDS-UPDRS-S scales.

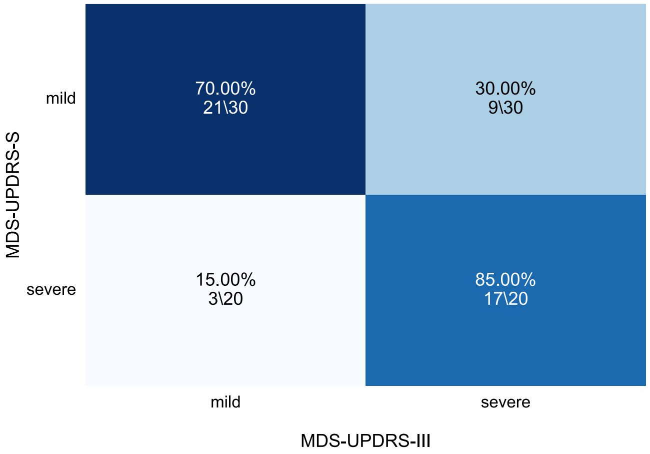


Fig. 5. Confusion matrix for the mild and severe classes labelled using the MDS-UPDRS-S and MDS-UPDRS-III ratings.

Table 2
Accuracy (in %) for classifying the PD severity level using the MDS-UPDRS-III scale. Accuracy is shown speaking task-wise (vowel task, sentence task, DDK task, read text task, and monologue task), feature-wise (phonation, articulation, prosody, and fusion), and classifier-wise (SVM, RF, MLP, and AdaBoost).

Speaking task	Feature	Classifier			
		SVM	RF	MLP	AdaBoost
Vowel task	Phonation	41 ± 0	39 ± 2	41 ± 1	47 ± 3
	Articulation	44 ± 2	46 ± 3	48 ± 4	47 ± 2
	Prosody	44 ± 2	36 ± 1	41 ± 3	38 ± 2
	Fusion	45 ± 1	48 ± 1	41 ± 1	47 ± 1
Sentence task	Phonation	36 ± 2	38 ± 1	35 ± 4	35 ± 1
	Articulation	45 ± 2	47 ± 5	46 ± 1	47 ± 5
	Prosody	41 ± 2	42 ± 2	45 ± 2	40 ± 6
	Fusion	47 ± 4	44 ± 6	43 ± 4	43 ± 7
DDK task	Phonation	39 ± 2	38 ± 1	38 ± 1	42 ± 0
	Articulation	50 ± 5	47 ± 6	44 ± 5	44 ± 4
	Prosody	42 ± 4	38 ± 3	38 ± 2	38 ± 2
	Fusion	46 ± 3	41 ± 5	37 ± 3	38 ± 3
Read text task	Phonation	46 ± 7	44 ± 4	43 ± 4	38 ± 3
	Articulation	47 ± 3	46 ± 6	50 ± 5	42 ± 3
	Prosody	43 ± 5	38 ± 3	53 ± 8	49 ± 3
	Fusion	46 ± 5	42 ± 7	49 ± 4	43 ± 5
Monologue task	Phonation	39 ± 2	51 ± 2	40 ± 1	53 ± 3
	Articulation	47 ± 3	46 ± 6	**58 ± 6**	42 ± 6
	Prosody	46 ± 2	42 ± 11	38 ± 3	46 ± 3
	Fusion	54 ± 3	40 ± 10	56 ± 2	48 ± 4

3.2. Evaluation metrics

Accuracy, which measures the correct predictions out of the total predictions performed, is used as the standard performance metric in the classification problem as in Pedregosa et al. (2011). In addition, confusion matrices are used to analyze how the classification systems function in the three classes. In this study, we used a balanced number of speakers per class. In each fold, the evaluation metrics were computed and stored. After the five folds, the final evaluation metrics were averaged.

4. Results

Tables 2 and 3 show the results of the severity level classification experiments using the MDS-UPDRS-III and MDS-UPDRS-S scales, respectively. The tables show the classification accuracy separately for all the three issues (speaking tasks, features, and classifiers) studied. For the MDS-UPDRS-III scale (Table 2), the highest accuracy of 58% was achieved for the monologue task using the articulation features with MLP as a classifier. For the MDS-UPDRS-S scale (Table 3), the best accuracy of 56% was achieved for the monologue task using the fusion features with MLP as a classifier. It can be seen from the tables that the accuracies were similar between the MDS-UPDRS-III and MDS-UPDRS-S scales.

The best classification systems reported in Table 2 for the MDS-UPDRS-III scale and in Table 3 for the MDS-UPDRS-S scale were further analyzed using confusion matrices and receiver operating characteristic (ROC) curves. Figs. 6 and 7 show the confusion matrix and the ROC curves, respectively, for the best system based on the MDS-UPDRS-III scale (i.e., the system utilizing the monologue task with the articulation features and the MLP classifier). The confusion matrix and ROC curves of the best system obtained for the MDS-UPDRS-S scale (i.e., the system using the monologue task with the fusion features and the MLP classifier) are shown in Figs. 8 and 9 respectively. For the confusion matrices, the actual and predicted labels are shown by the vertical and horizontal axes, respectively. For the ROC curves, the true positive rate (TPR) and false positive rate (FPR) are shown by the vertical and horizontal axes, respectively. The ROC curves are plotted for each of the classes separately as one versus all. The micro-averaging ROC curve was obtained by evaluating every element of the class indicator matrix as a binary classification problem. The macro-averaging ROC curve assigns similar weights to the classification of each class. Fig. 6 demonstrates that for the MDS-UPDRS-III scale, healthy and severe classes were classified better than the mild class. It can be seen that the healthy class was often incorrectly classified as mild and the severe class also as mild. From the ROC curves (Fig. 7), it can be seen that mild was the hardest class to detect, and the healthy and severe classes were easiest to distinguish. Similar observations can also be observed from the confusion matrix. From Fig. 8, it can be observed that for the MDS-UPDRS-S scale, the mild and healthy classes were classified better than the severe class. The figure shows that also for the MDS-UPDRS-S scale, the severe class was often mixed with the mild class. From the ROC curves (Fig. 9), it can be observed that mild and severe classes were the hardest classes to detect, and healthy was easiest to distinguish. It is important to note that the lack of MDS-UPDRS scores for the HCs may have impacted the ability to distinguish speech of the PD patients with mild symptoms from speech of the HCs.

The mean accuracy values reported in Tables 2 and 3 show, in general, minor differences between the speaking tasks, features, and classifiers for both of the MDS-UPDRS scales. A further statistical evaluation was conducted to analyze whether the data shows

Table 3
Accuracy (in %) for classifying the PD severity level using the MDS-UPDRS-S scale. Accuracy is shown speaking task-wise (vowel task, sentence task, DDK task, read text task, and monologue task), feature-wise (phonation, articulation, prosody, and fusion), and classifier-wise (SVM, RF, MLP, and AdaBoost).

Speaking task	Feature	Classifier			
		SVM	RF	MLP	AdaBoost
Vowel task	Phonation	51 ± 5	52 ± 3	53 ± 5	45 ± 4
	Articulation	50 ± 2	50 ± 3	53 ± 0	52 ± 2
	Prosody	42 ± 3	39 ± 2	36 ± 4	40 ± 3
	Fusion	49 ± 3	55 ± 3	48 ± 1	55 ± 2
Sentence task	Phonation	43 ± 4	41 ± 2	41 ± 1	42 ± 1
	Articulation	42 ± 5	46 ± 6	44 ± 4	40 ± 5
	Prosody	46 ± 4	47 ± 2	46 ± 2	47 ± 2
	Fusion	46 ± 5	47 ± 1	49 ± 2	45 ± 4
DDK task	Phonation	41 ± 2	45 ± 3	40 ± 1	45 ± 2
	Articulation	40 ± 5	51 ± 7	42 ± 4	48 ± 6
	Prosody	42 ± 3	45 ± 3	47 ± 5	48 ± 4
	Fusion	48 ± 4	43 ± 4	47 ± 1	39 ± 2
Read text task	Phonation	50 ± 4	37 ± 5	50 ± 3	37 ± 2
	Articulation	39 ± 2	37 ± 2	35 ± 3	35 ± 2
	Prosody	37 ± 2	39 ± 4	41 ± 2	48 ± 2
	Fusion	37 ± 4	39 ± 9	31 ± 5	41 ± 5
Monologue task	Phonation	41 ± 2	48 ± 10	37 ± 8	44 ± 2
	Articulation	46 ± 1	44 ± 3	50 ± 3	44 ± 3
	Prosody	46 ± 5	49 ± 8	39 ± 1	52 ± 0
	Fusion	44 ± 2	44 ± 3	**56 ± 4**	50 ± 2

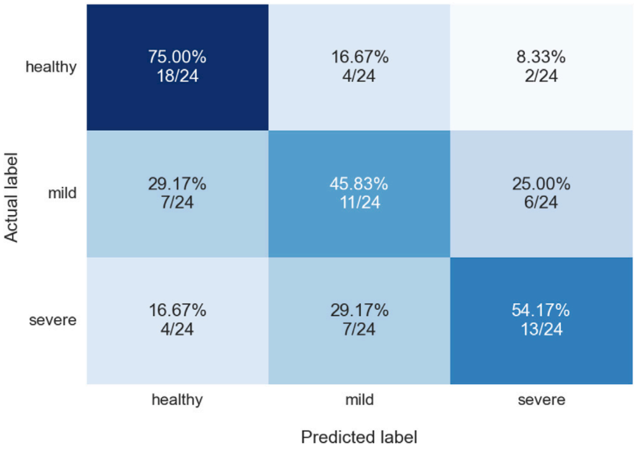


Fig. 6. Confusion matrix of the best system (the monologue task, articulation features, MLP classifier) obtained using the MDS-UPDRS-III scale for the severity level classification task.

significant differences between the different speaking tasks, features, and classifiers. We used Cochran's Q test (Fleiss et al., 1981; Kuncheva, 2014; Cochran, 1950; Raschka, 2018), a generalized version of McNemar's test, in the statistical evaluation. Cochran's Q test can be used to evaluate multiple classification systems, and it has been used previously in similar studies (Narendra and Alku, 2021). In its basic form, Cochran's Q test compares binary classification systems. In our study, the test was used for comparing 3-level classifiers by marking their output as 1 or 0 depending on whether the system corrected classified the underlying class (output = 1) or incorrectly classified it as one of the other two classes (output = 0).

By using a significance level of alpha = 0.05, three Cochran's Q tests were first conducted for the accuracy values of both of the MDS-UPDRS scales by considering the speaking task, feature, and classifier as independent variables. For the MDS-UPDRS-III scale, these tests indicated that the null hypothesis (i.e., no difference between the classification accuracies) was rejected for two variables (the speaking task and feature). Multiple post hoc pair-wise tests were then conducted for the different speaking tasks and features using the McNemar test with Bonferroni correction. Regarding the speaking tasks, these post hoc tests indicated that the monologue task was significantly better compared to the vowel task (χ^2 = 29.34, $p <$ 1.0e−07), sentence task (χ^2 = 35.28, $p <$ 1.0e−08), and DDK task (χ^2 = 41.30, $p <$ 1.0e−09). In addition, the read text task was also significantly better than the vowel task (χ^2 = 19.56, $p <$ 1.0e−05), the sentence task (χ^2 = 24.35, $p <$ 1.0e−06), and the DDK task (χ^2 = 29.93, $p <$ 1.0e−07). Furthermore, these post hoc tests indicated that the articulation features were significantly better than the phonation (χ^2 = 56.04, $p <$ 1.0e−13)

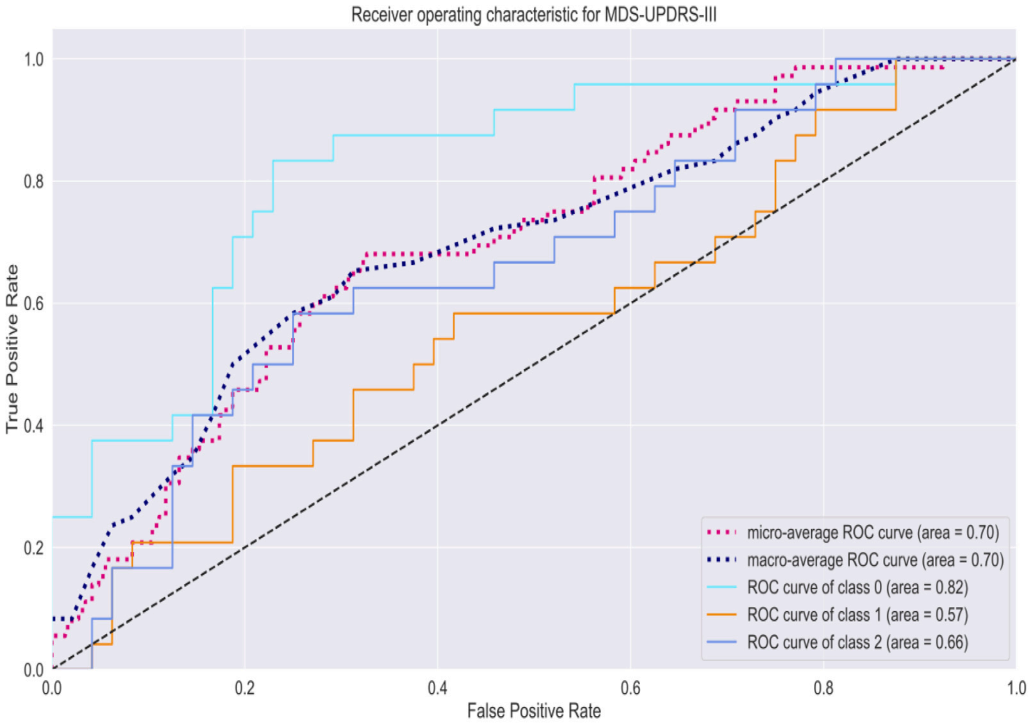


Fig. 7. Multi-class ROC curve of the best system (the monologue task, articulation features, MLP classifier) obtained using the MDS-UPDRS-III scale. Class 0: healthy, Class 1: mild, Class 2: severe.

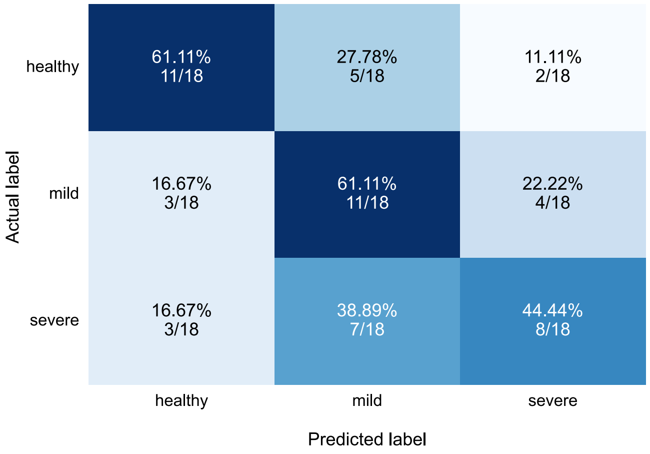


Fig. 8. Confusion matrix of the best system (the monologue task, fusion features, MLP classifier) obtained using the MDS-UPDRS-S scale for the severity level classification task.

and prosody (χ^2 = 50.85, p < 1.0e−12) features. Moreover, the fusion features were significantly better than the phonation (χ^2 = 42.06, p < 1.0e−10) and prosody (χ^2 = 41.04, p < 1.0e−09) features.

For the MDS-UPDRS-S scale, Cochran's Q test indicated that the null hypothesis was rejected for all three issues (speaking task, feature, and classifier). Multiple post hoc pair-wise tests were then conducted for the different speaking tasks, features, and classifiers. Regarding the speaking tasks, these post hoc tests indicated that the monologue task was significantly better compared to the vowel task (χ^2 = 7.34, p < 1.0e−02), the sentence task (χ^2 = 9.69, p < 1.0e−02), and the DDK task (χ^2 = 23.56, p < 1.0e−05). In addition, the read text task was significantly better than the DDK task (χ^2 = 11.23, p < 1.0e−03), and the sentence task was also better than the DDK task (χ^2 = 4.25, p < 1.0e−02). Furthermore, these post hoc tests indicated that the articulation features were significantly better than the phonation (χ^2 = 6.84, p < 1.0e−02) and prosody (χ^2 = 39.75, p < 1.0e−09) features. Moreover, the fusion features were significantly better than the phonation (χ^2 = 15.18, p < 1.0e−04) and prosody (χ^2 = 66.30, p < 1.0e−15) features. The

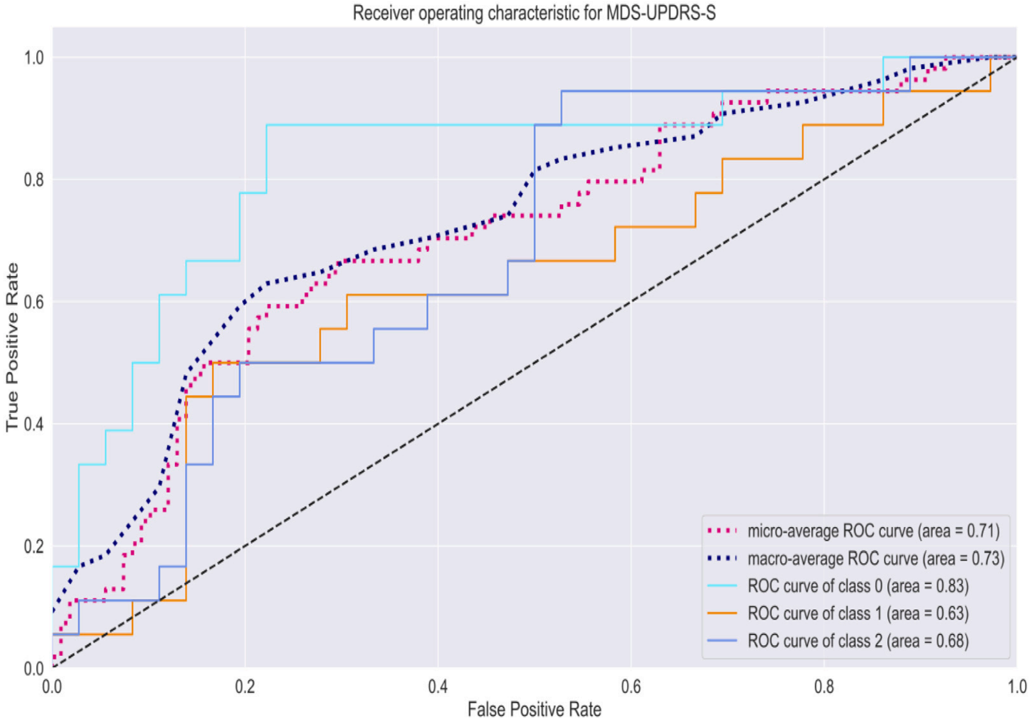


Fig. 9. Multi-class ROC curve of the best system (the monologue task, fusion features, MLP classifier) obtained using the MDS-UPDRS-S scale. Class 0: healthy, Class 1: mild, Class 2: severe.

phonation features were also significantly better than the prosody (χ^2 = 13.85, p < 1.0e−03) features. Finally, these post hoc tests indicated that the MLP classifier was significantly better than AdaBoost (χ^2 = 4.47, p < 1.0e−02).

5. Discussion and conclusions

The automatic classification of PD from speech has been studied in many investigations during the past two decades. However, compared to this binary classification problem, much less research has been devoted to the automatic multi-class classification of the severity level of PD from speech. This ML problem has been investigated in only a few studies (Arias-Londoño and Gómez-García, 2019; Arias-Vergara et al., 2018a; Vásquez-Correa et al., 2018a) by using few speaking tasks, features, and classifiers. Therefore, the goal of the present study was to conduct a systematic comparison between the different speaking tasks, features, and classifiers that have been used in speech-based biomarking studies in PD. For these three issues, the current study investigated five speaking tasks (vowel task, sentence task, DDK task, read text task, and monologue task), four features (phonation, articulation, prosody, and their fusion), and four classifier architectures (SVM, RF, MLP, and AdaBoost). The studied ML task was the automatic classification of speech into three PD severity level classes (healthy *vs*. mild PD *vs*. severe PD). For the ground truth severity level labels, two MDS-UPDRS scales (MDS-UPDRS-III and the MDS-UPDRS-S) were used.

In the classification of the PD severity level based on the MDS-UPDRS-III scale, the accuracy obtained by using individual combinations of speaking tasks, features, and classifiers varied from 32% and 58%. The highest accuracy was obtained for the combination of the monologue task, the articulation features and the MLP classifier. For the MDS-UPDRS-S scale, the accuracy varied between 31% and 56%, and the highest value was obtained by using the monologue task, the fusion features and the MLP classifier. In general, the differences in accuracy between the MDS-UPDRS-III and the MDS-UPDRS-S scales were small. We consider this result interesting because one could have expected that the speech-based classification based on the MDS-UPDRS-S scale is an easier task for the computer – and the accuracy values would therefore be higher than for the MDS-UPDRS-III scale – because the severity level classes to be predicted are based on ground truth labels which are also speech-based (i.e., the MDS-UPDRS-S scale uses just one modality, speech, in PD severity level assessment (Goetz et al., 2008)). However, the smallish difference in the accuracy values between the MDS-UPDRS-III and MDS-UPDRS-S scales suggests that the speech-based automatic classification is not affected by the modalities used in the PD severity level assessment to define the ground truth labels. By comparing the current results with those reported in similar previous studies, it can be noted that the best accuracy of 58% obtained using the MDS-UPDRS-III scale in the current study was clearly better than the best accuracy of 51% reported (Arias-Londoño and Gómez-García, 2019). For the MDS-UPDRS-S scale, too, the best accuracy of 56% obtained in this study was better compared to the best accuracy of 54.9% reported in Vásquez-Correa et al. (2018a). However, direct comparisons with previous studies should be treated as approximately because many factors (e.g., database, cross-validation, etc.) differ between the previous studies and the current one. Nevertheless, the current results are in line with the results of previous studies and indicate that the studied 3-class classification problem is challenging for every combination of the studied speaking tasks, features, and classifiers because the accuracy of even the best system was only 58%, which is less than twice the chance level (of 33.33%). Despite the achieved accuracies may not be high enough for practical applications, we argue that our study offers valuable insights into comparing the performance of various speaking tasks, features, and classifiers for 3-class PD severity classification. We hope that the findings reported in this study can guide future research to develop classification models that are more accurate and therefore reliable also for practical applications.

Statistical tests were conducted for both MDS-UPDRS scales to analyze whether classification accuracy shows significant differences between the different speaking tasks, features, and classifiers. For the speaking tasks, the tests revealed that the monologue and the read text tasks were significantly better than the three other tasks (the vowel task, the sentence task, and the DDK task) for the MDS-UPDRS-III scale. However, for the MDS-UPDRS-S scale, the monologue task was better than the vowel, sentence, and DDK tasks, whereas the read text task showed a significant improvement only compared to the DDK task. A significant improvement was shown by the monologue task, which captures the complexity of speech production with a combination of speech motor execution and cognitive-linguistic processing, both of which are impaired in PD patients (Eyigoz et al., 2020; Šimek and Rusz, 2021). Although no cognitive impairment assessment information is provided in the PC-GITA corpus, it is plausible that cognitive impairment may also have contributed to the better performance of the monologue task compared to the other speaking tasks (García et al., 2021; Rusz and Tykalová, 2021). Furthermore, the length of the speaking task may have also influenced the results, as demonstrated by recent research showing that longer monologues lead to better stability of acoustic outcomes (Krỳže et al., 2021). Additionally, previous studies have demonstrated that PD patients are typically more intelligible in prepared utterances compared to spontaneous speech (Kim et al., 2011; Kempler and Van Lancker, 2002). In Kempler and Van Lancker (2002), both spontaneous and non-spontaneous speech were collected from a single PD patient who was diagnosed 18 years before the investigation, and the intelligibility of spontaneous speech was found to be severely affected when compared to non-spontaneous speech. In Rusz et al. (2013), authors analyzed vowel articulation across various speaking tasks, including sustained phonations, sentence repetitions, text reading, and reading a monologue, in a group of 20 early PD patients and 15 HCs. The authors used a set of features that consisted of measures of the first and second formants (F1 and F2), the vowel articulation index (VAI), the vowel space area (VSA), and the F2i/F2u quotient (the ratio between F2 in /i/ and F2 in /u/). The findings revealed that sustained phonations were not suitable for evaluating vowel articulation in PD patients, while the monologue task was the most sensitive task for differentiating between PD patients and HCs. The authors were able to achieve classification scores of around 80% by applying different articulation measures (such as VSA and F2i/F2u) to the monologue, demonstrating the potential to distinguish between PD patients and HCs. Our results are in line with the findings of previous studies (Kempler and Van Lancker, 2002; Rusz et al., 2013; Šimek and Rusz, 2021), which suggest that the monologue task is the most affected speaking task in classifying the severity of PD. These results suggest that the monologue task is recommended for collecting speech data to study the classification studies of the PD severity level. From the feature sets studied, the articulation and fusion features were significantly better than the phonation and prosody features. The better performance of the articulation and fusion features is most likely due to the involvement of such individual features in these two sets that properly reflect the effects caused by moving articulators such as the tongue, jaw and lips, which are known to be affected by PD. With respect to the classifiers, the statistical tests indicated only one significant difference in the pair-wise comparisons (i.e., MLP was better than AdaBoost for the MDS-UPDRS-S scale). Hence, compared to the speaking tasks and features, the selection of the classifier architecture clearly has a smaller effect on the accuracy of the PD severity level classification system.

In conclusion, a systematic investigation was conducted to study three issues (speaking tasks, features, classifiers) in the automatic classification of the severity level of PD into three classes using the speech signal. The results showed that the use of the monologue task and the articulation and fusion features improved classification accuracy significantly compared to the other speaking tasks and features. In addition, the accuracy differences between the systems based on the MDS-UPDRS-III and MDS-UPDRS-S scales were generally small. However, the studied 3-class ML problem turned out to be challenging using any of the compared systems, and the best combination of the three issues resulted in a mediocre accuracy of roughly 58%. It should be noted that our results are all based on experiments conducted using a single, smallish database because we did not have access to larger open repositories of parkinsonian speech. To develop the automatic severity level classification technology for clinical use, larger databases are definitely needed. In order to improve classification performance, future research should focus more on, for example, deep learning-based end-to-end systems (Narendra and Alku, 2021) that are not based on the similar classical two-stage structure, which was studied in this investigation. This, however, further increases the importance of larger, MDS-UPDRS-labelled speech databases to get adequate amounts of data to train data-hungry deep learning-based architectures. Finally, it is worth noting that PD can manifest itself in various forms for different patients. While voice disorders are known to be very common in individuals with PD (the prevalence of voice disorders in PD patients has been reported to vary between 70% and 90% (Logemann et al., 1978; Hartelius and Svensson, 1994; Ho et al., 1998; Ma et al., 2020)), the absence of voice impairments in some PD patients should be recognized as a potential limitation in the study.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data was obtained from JR.Orozco-Arroyave.

Acknowledgements

This study was funded by the Academy of Finland, Finland (project no. 330139) and Aalto University (the Ministry of Education and Culture program for India). Aalto Science IT provided the computational resources.

References

Arias-Londoño, Julián D., Gómez-García, Jorge A., 2019. Predicting UPDRS scores in Parkinson's disease using voice signals: A deep learning/transfer-learning-based approach. In: Automatic Assessment of Parkinsonian Speech Workshop. Springer, pp. 100–123.

Arias-Vergara, Tomas, Vásquez-Correa, Juan Camilo, Orozco-Arroyave, Juan Rafael, 2017. Parkinson's disease and aging: analysis of their effect in phonation and articulation of speech. Cogn. Comput. 9 (6), 731–748.

Arias-Vergara, Tomas, Vasquez-Correa, Juan Camilo, Orozco-Arroyave, Juan Rafael, Klumpp, Philipp, Nöth, Elmar, 2018a. Unobtrusive monitoring of speech impairments of Parkinson's disease patients through mobile devices. In: International Conference on Acoustics, Speech and Signal Processing. pp. 6004–6008.

Arias-Vergara, Tomas, Vasquez-Correa, Juan Camilo, Orozco-Arroyave, Juan Rafael, Nöth, Elmar, 2018b. Speaker models for monitoring Parkinson's disease progression considering different communication channels and acoustic conditions. Speech Commun. 101, 11–25.

Bocklet, Tobias, Steidl, Stefan, Nöth, Elmar, Skodda, Sabine, 2013. Automatic evaluation of Parkinson's speech-acoustic, prosodic and voice related cues. In: Interspeech. pp. 1149–1153.

Cernak, Milos, Orozco-Arroyave, Juan Rafael, Rudzicz, Frank, Christensen, Heidi, Vásquez-Correa, Juan Camilo, Nöth, Elmar, 2017. Characterisation of voice quality of Parkinson's disease using differential phonological posterior features. Comput. Speech Lang. 46, 196–208.

Cochran, William G., 1950. The comparison of percentages in matched samples. Biometrika 37 (3/4), 256–266.

Cortes, Corinna, Vapnik, Vladimir, 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.

Dehak, Najim, Dumouchel, Pierre, Kenny, Patrick, 2007. Modeling prosodic features with joint factor analysis for speaker verification. IEEE Trans. Audio, Speech, Lang. Process. 15 (7), 2095–2103.

Erdogdu Sakar, Betul, Serbes, Gorkem, Sakar, C. Okan, 2017. Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease. PLoS One 12 (8), e0182428.

Eyigoz, Elif, Courson, Melody, Sedeño, Lucas, Rogg, Katharina, Orozco-Arroyave, Juan Rafael, Nöth, Elmar, Skodda, Sabine, Trujillo, Natalia, Rodríguez, Mabel, Rusz, Jan, et al., 2020. From discourse to pathology: automatic identification of Parkinson's disease patients via morphological measures across three languages. Cortex 132, 191–205.

Falk, Tiago H., Chan, Wai-Yip, Shein, Fraser, 2012. Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. Speech Commun. 54 (5), 622–631.

Fleiss, Joseph L., Levin, Bruce, Paik, Myunghee Cho, 1981. Statistical Methods for Rates and Proportions, second ed. John Wiley & Sons, New York.

García, Adolfo M, Arias-Vergara, Tomás, C Vasquez-Correa, Juan, Nöth, Elmar, Schuster, Maria, Welch, Ariane E, Bocanegra, Yamile, Baena, Ana, Orozco-Arroyave, Juan R, 2021. Cognitive determinants of dysarthria in Parkinson's disease: an automated machine learning approach. Mov. Disorders 36 (12), 2862–2873.

Garcia, Nicanor, Orozco-Arroyave, Juan Rafael, Luis Fernando, D'Haro, Dehak, Najim, Nöth, Elmar, 2017. Evaluation of the neurological state of people with Parkinson's disease using i-vectors. In: Interspeech. pp. 299–303.

Goetz, Christopher G, Tilley, Barbara C, Shaftman, Stephanie R, Stebbins, Glenn T, Fahn, Stanley, Martinez-Martin, Pablo, Poewe, Werner, Sampaio, Cristina, Stern, Matthew B, Dodel, Richard, et al., 2008. Movement disorder society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. Mov. Disorders: Official J. Mov. Disorder Soc. 23 (15), 2129–2170.

Gómez-Vilda, Pedro, Gómez-Rodellar, Andrés, Palacios-Alonso, Daniel, Rodellar-Biarge, Victoria, Álvarez-Marquina, Agustín, 2022. The role of data analytics in the assessment of pathological speech—A critical appraisal. Appl. Sci. 12 (21), 11095.

Hartelius, Lena, Svensson, Per, 1994. Speech and swallowing symptoms associated with Parkinson's disease and multiple sclerosis: A survey. Folia Phoniatrica Logopaedica 46, 9–17.

Hegde, Sarika, Shetty, Surendra, Rai, Smitha, Dodderi, Thejaswi, 2019. A survey on machine learning approaches for automatic detection of voice disorders. J. Voice 33 (6), 947–e11.

Hlavnička, Jan, Čmejla, Roman, Tykalová, Tereza, Šonka, Karel, Růžička, Evžen, Rusz, Jan, 2017. Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder. Sci. Rep. 7 (1), 1–13.

Ho, Aileen K, Iansek, Robert, Marigliani, Caterina, Bradshaw, John L, Gates, Sandra, 1998. Speech impairment in a large sample of patients with Parkinson's disease. Behav. Neurol. 11, 131–137.

Hornykiewicz, Oleh, 1998. Biochemical aspects of Parkinson's disease. Neurology 51 (2 Suppl 2), S2–S9.

Jankovic, Joseph, 2008. Parkinson's disease: clinical features and diagnosis. J. Neurol., Neurosurg. Psychiatry 79 (4), 368–376.

Kadiri, Sudarsana Reddy, Kethireddy, Rashmi, Alku, Paavo, 2020. Parkinson's disease detection from speech using single frequency filtering cepstral coefficients. In: Interspeech. pp. 4971–4975.

Karan, Biswajit, Sahu, Sitanshu Sekhar, Orozco-Arroyave, Juan Rafael, Mahto, Kartik, 2021. Non-negative matrix factorization-based time-frequency feature extraction of voice signal for Parkinson's disease prediction. Comput. Speech Lang. 69, 101216.

Kempler, Daniel, Van Lancker, Diana, 2002. Effect of speech task on intelligibility in dysarthria: A case study of Parkinson's disease. Brain Lang. 80 (3), 449–464.

Kim, Yunjung, Kent, Raymond D., Weismerb, Gary, 2011. An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria. J. Speech, Lang., Hear. Res. 54, 417–429.

Krỳže, Petr, Tykalová, Tereza, Ržička, Evžen, Rusz, Jan, 2021. Effect of reading passage length on quantitative acoustic speech assessment in Czech-speaking individuals with Parkinson's disease treated with subthalamic nucleus deep brain stimulation. J. Acoust. Soc. Am. 149 (5), 3366–3374.

Kuncheva, Ludmila I., 2014. Combining Pattern Classifiers: Methods and Algorithms. John Wiley & Sons, Canada.

Logemann, Jeri A, Fisher, Hilda B, Boshes, Benjamin, Blonsky, E Richard, 1978. Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. J. Speech Hear. Disorders 43 (1), 47–57.

López, José Vicente Egas, Orozco-Arroyave, Juan Rafael, Gosztolya, Gábor, 2019. Assessing Parkinson's disease from speech using Fisher vectors. In: Interspeech.

Ma, Andrew, Lau, Kenneth K., Thyagarajan, Dominic, 2020. Voice changes in Parkinson's disease: What are they telling us? J. Clin. Neurosci. 72, 1–7.

Moro-Velazquez, Laureano, Gomez-Garcia, Jorge A, Godino-Llorente, Juan I, Grandas-Perez, Francisco, Shattuck-Hufnagel, Stefanie, Yagüe-Jimenez, Virginia, Dehak, Najim, 2019a. Phonetic relevance and phonemic grouping of speech in the automatic detection of Parkinson's disease. Sci. Rep. 9 (1), 1–16.

Moro-Velazquez, Laureano, Gomez-Garcia, Jorge Andres, Godino-Llorente, Juan Ignacio, Villalba, Jesús, Rusz, Jan, Shattuck-Hufnagel, Stephanie, Dehak, Najim, 2019b. A forced Gaussians based methodology for the differential evaluation of Parkinson's disease by means of speech processing. Biomed. Signal Process. Control 48, 205–220.

Narendra, N.P., Airaksinen, Manu, Story, Brad, Alku, Paavo, 2019. Estimation of the glottal source from coded telephone speech using deep neural networks. Speech Commun. 106, 95–104.

Narendra, N.P., Alku, Paavo, 2021. Automatic assessment of intelligibility in speakers with dysarthria from coded telephone speech using glottal features. Comput. Speech Lang. 65, 101117.

Nilashi, Mehrbakhsh, Ibrahim, Othman, Samad, Sarminah, Ahmadi, Hossein, Shahmoradi, Leila, Akbari, Elnaz, 2019. An analytical method for measuring the Parkinson's disease progression: A case on a Parkinson's telemonitoring dataset. Measurement 136, 545–557.

Nussbaum, Robert L., Ellis, Christopher E., 2003. Alzheimer's disease and Parkinson's disease. N. Engl. J. Med. 348 (14), 1356–1364.

Oktay, Ayse Betul, Kocer, Abdulkadir, 2020. Differential diagnosis of Parkinson and essential tremor with convolutional LSTM networks. Biomed. Signal Process. Control 56, 101683.

Orozco-Arroyave, Juan Rafael, Arias-Londoño, Julián David, Vargas-Bonilla, Jesús Francisco, Gonzalez-Rátiva, María Claudia, Nöth, Elmar, 2014. New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 342–347.

Orozco-Arroyave, Juan Rafael, Hönig, F, Arias-Londoño, JD, Vargas-Bonilla, JF, Daqrouq, K, Skodda, S, Rusz, J, Nöth, E, 2016. Automatic detection of Parkinson's disease in running speech spoken in three different languages. J. Acoust. Soc. Am. 139 (1), 481–500.

Orozco-Arroyave, Juan Rafael, Vásquez-Correa, Juan Camilo, Vargas-Bonilla, Jesús Francisco, Arora, Raman, Dehak, Najim, Nidadavolu, Phani S, Christensen, Heidi, Rudzicz, Frank, Yancheva, Maria, Chinaei, H, 2018. Neurospeech: an open-source software for Parkinson's speech analysis. Digit. Signal Process. 77, 207–221.

Pal, Mahesh, 2005. Random forest classifier for sensing classification. Int. J. Remote Sens. 26 (1), 217–222.

Pal, Sankar K., Mitra, Sushmita, 1992. Multilayer perceptron, fuzzy sets, and classification. IEEE Trans. Neural Netw. 3 5, 683–697.

Parisi, Luca, RaviChandran, Narrendar, Manaog, Marianne Lyne, 2018. Feature-driven machine learning to improve early diagnosis of Parkinson's disease. Expert Syst. Appl. 110, 182–190.

Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent, 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pérez-Toro, Paula Andrea, Vásquez-Correa, Juan Camilo, Strauss, M, Orozco-Arroyave, Juan Rafael, Nöth, Elmar, 2019. Natural language analysis to detect Parkinson's disease. In: International Conference on Text, Speech, and Dialogue. pp. 82–90.

Raschka, Sebastian, 2018. Mlxtend: providing machine learning and data science utilities and extensions to Python's scientific computing stack. J. Open Source Softw. 3 (24).

Rusz, Jan, Cmejla, Roman, Ruzickova, Hana, Ruzicka, Evzen, 2011. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. J. Acoust. Soc. Am. 129 (1), 350–367.

Rusz, Jan, Cmejla, Roman, Tykalova, Tereza, Ruzickova, Hana, Klempir, Jiri, Majerova, Veronika, Picmausova, Jana, Roth, Jan, Ruzicka, Evzen, 2013. Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task. J. Acoust. Soc. Am. 134 (3), 2171–2181.

Rusz, Jan, Hlavnička, Jan, Novotnỳ, Michal, Tykalová, Tereza, Pelletier, Amelie, Montplaisir, Jacques, Gagnon, Jean-Francois, Dušek, Petr, Galbiati, Andrea, Marelli, Sara, et al., 2021a. Speech biomarkers in rapid eye movement sleep behavior disorder and Parkinson disease. Ann. Neurol. 90 (1), 62–75.

Rusz, Jan, Tykalová, Tereza, 2021. Does cognitive impairment influence motor speech performance in de novo Parkinson's disease? Mov. Disorders 36 (12), 2980–2982.

Rusz, Jan, Tykalová, Tereza, Novotnỳ, Michal, Ržička, Evžen, Dušek, Petr, 2021b. Distinct patterns of speech disorder in early-onset and late-onset de-novo Parkinson's disease. npj Parkinson's Dis. 7 (1), 98.

Rusz, Jan, Tykalová, Tereza, Novotnỳ, Michal, Zogala, David, Ržička, Evžen, Dušek, Petr, 2022. Automated speech analysis in early untreated Parkinson's disease: relation to gender and dopaminergic transporter imaging. Eur. J. Neurol. 29 (1), 81–90.

Schapire, Robert E., 2013. Explaining Adaboost. In: Empirical Inference. pp. 37–52.

Shahbakhi, Mohammad, Far, Danial Taheri, Tahami, Ehsan, 2014. Speech analysis for diagnosis of Parkinson's disease using genetic algorithm and support vector machine. J. Biomed. Sci. Eng. 7 (4), 147–156.

Shinde, Sumeet, Prasad, Shweta, Saboo, Yash, Kaushick, Rishabh, Saini, Jitender, Pal, Pramod Kumar, Ingalhalikar, Madhura, 2019. Predictive markers for Parkinson's disease using deep neural nets on neuromelanin sensitive MRI. NeuroImage: Clin. 22, 101748.

Šimek, Michal, Rusz, Jan, 2021. Validation of cepstral peak prominence in assessing early voice changes of Parkinson's disease: Effect of speaking task and ambient noise. J. Acoust. Soc. Am. 150 (6), 4522–4533.

Tanner, Caroline M., Goldman, Samuel M., 1996. Epidemiology of Parkinson's disease. Neurol. Clin. 14 (2), 317–335.

Vásquez-Correa, Juan Camilo, Arias-Vergara, Tomas, Orozco-Arroyave, Juan Rafael, Eskofier, Björn, Klucken, Jochen, Nöth, Elmar, 2018a. Multimodal assessment of Parkinson's disease: A deep learning approach. IEEE J. Biomed. Health Inf. 23 (4), 1618–1630.

Vásquez-Correa, Juan Camilo, Orozco-Arroyave, JR, Bocklet, T, Nöth, E, 2018b. Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease. J. Commun. Disorders 76, 21–36.

Vásquez-Correa, Juan Camilo, Orozco-Arroyave, Juan Rafael, Nöth, Elmar, 2017. Convolutional neural network to model articulation impairments in patients with Parkinson's disease. In: Interspeech. pp. 314–318.

Verde, Laura, De Pietro, Giuseppe, Sannino, Giovanna, 2018. Voice disorder identification by using machine learning techniques. IEEE Access 6, 16246–16255.

Wang, Meng, Wen, Yanxia, Mo, Shicong, Yang, Liqiong, Chen, Xiaqing, Luo, Man, Yu, Hongdian, Xu, Fan, Zou, Xianwei, 2022. Distinctive acoustic changes in speech in Parkinson's disease. Comput. Speech Lang. 75, 101384.