



FACULTAT  
**DE CIÈNCIES  
I TECNOLOGIA**

**UVIC** | UVIC·UCC

## Degree Final Project

# Acoustic Analysis of Speech and Voice Disorders in Patients with Lewy Body Diseases

---

Andrea Fernández Martínez

**Degree in Biomedical Engineering**

Tutor: Jiri Mekyska

Endorser: Jordi Solé Casals

Vic, June 2019



# Acknowledgements

---

First of all, I would like to acknowledge my family for their support through all these years. Without them, nothing would have been possible. Thanks to my parents, my grandparents and my brother for always being by my side. I am who I am thanks to you.

I would also like to acknowledge the Brain Disease Analysis Laboratory, Department of Telecommunications (Brno University of Technology, Czech Republic) for hosting me, supporting me and giving me the opportunity to grow personally and professionally. I will always remember the great semester I spend in Czech Republic and everything I have learned there. Special appreciation to my supervisor Jiri Mekyska for all the help and support, this would have not be possible without you.

To the UVic – UCC for all these years in university. Thank you for the high quality education you provide and for the always good treatment, you have been a second family during these years. To my supervisor in UVic - UCC, Jordi Solé, for all the help and advises you have provided me with. Thank you for believing in me and the support you have always shown.

To SeidorLabs, thank you for giving me an opportunity and for supporting young students. To my friends, to all the people that have accompanied me these past 4 years, thank you for being part of this amazing journey. To the people I have met during my ERASMUS+, thank you for showing me another world, I will be forever thankful for all the things I have seen and learned from you.

Thank you.

# Index

---

Abstract .....	5
Resumen .....	6
Introduction .....	7
Motivation .....	8
Objectives .....	8
State of the art .....	9
Lewy Bodies Diseases .....	9
Parkinson's disease.....	9
Lewy Bodies dementias: Parkinson's disease with Dementia and Dementia with Lewy Bodies .	10
Current research in Lewy Bodies Diseases.....	11
Theoretical Fundamentals .....	12
Introduction to Machine Learning.....	12
Evaluation of the performance of ML algorithms .....	12
Data analysis and Machine Learning algorithms.....	14
Confidence intervals .....	14
Feature selection approaches .....	15
Logistic Regression.....	16
XGBoost.....	16
Methodology.....	18
Participants .....	18
Acoustic analysis.....	18
Statistical Analysis .....	22
Univariate analysis.....	22
Multivariate analysis.....	22
Results .....	23
Discussion .....	30
Conclusion.....	31
Project acknowledgments .....	32
References .....	32

# Abstract

---

Lewy body diseases (LBDs) are a group of neurodegenerative diseases that consists of Parkinson's disease, Parkinson's disease with dementia (PD), and dementia with Lewy bodies. LBDs are characterized by the aggregation of  $\alpha$ -synuclein in specific brain regions, which leads to the formation of Lewy bodies and Lewy neuritis, usually accompanied by neurodegeneration. Identification of early stages in LBDs is crucial since the neurodegeneration may be possibly stopped or treated before the pathological cascades start.

This study aims to identify people at high risk of developing LBDs i.e. in prodromal stages of these diseases, based on an acoustic analysis of speech and voice. Eighteen acoustic features were evaluated from 36 Czech native, clinically diagnosed PD patients, 29 healthy controls (HC) and 37 subjects at risk of developing LBDs. Statistical analysis of acoustic features and multiple modelling techniques were explored in order to find the best discriminative model, including multivariate logistic regression and the state-of-the-art algorithm XGBoost.

We finally propose an XGBoost model that identifies subjects in high risk of developing LBDs achieving 76 % specificity and 71 % sensitivity on test dataset (including PD and HC). Prosodic and articulatory features were found to be the most discriminant. Thirteen subjects were confirmed to be in high risk of LBDs with our model. Accuracy of these results will be further evaluated in future following a longitudinally monitoring and clinical examination of participants.

## Resumen

---

Las enfermedades de cuerpos de Lewy son un grupo de enfermedades neurodegenerativas, que incluye la enfermedad de Parkinson, la enfermedad de Parkinson con demencia y demencia con cuerpos de Lewy. Este grupo de enfermedades se caracteriza por la agregación de  $\alpha$ -sinucleína en regiones específicas del cerebro, lo cual da lugar a la formación de cuerpos de Lewy y neuritas de Lewy, normalmente acompañado de neurodegeneración. La identificación de los estados iniciales de estas enfermedades es crucial, ya que la neurodegeneración puede ser potencialmente parada o tratada antes de que las cascadas patológicas comiencen.

Este estudio tiene como objetivo identificar sujetos en alto riesgo de desarrollar enfermedades de cuerpos de Lewy mediante un análisis acústico de voz y habla. Dieciocho características acústicas han sido evaluadas de 36 pacientes de nacionalidad checa clínicamente diagnosticados con Parkinson, 29 controles sanos y 37 individuos en riesgo de desarrollar enfermedades de cuerpos de Lewy. Se han explorado diferentes análisis estadísticos y múltiples técnicas de modelaje para identificar el mejor modelo discriminativo, incluyendo regresiones logísticas multivariantes y el algoritmo state-of-the-art XGBoost.

Finalmente, se ha propuesto un modelo XGBoost que identifica a sujetos en riesgo de desarrollar enfermedades de cuerpos de Lewy con 76 % de especificidad y 71 % de sensibilidad en el conjunto de datos test (incluyendo pacientes de Parkinson y controles sanos). Las características más discriminativas halladas han sido prosódicas y articulatorias. Trece sujetos han sido confirmados de estar en alto riesgo de desarrollar este tipo de enfermedades con nuestro modelo. La precisión de estos resultados será evaluada en el futuro siguiendo un monitoreo longitudinal y exámenes clínicos de los pacientes.

# Introduction

---

Lewy body diseases (LBDs) are a group of neurodegenerative diseases that consists of Parkinson's disease (PD), Parkinson's disease with Dementia (PDD) and Dementia with Lewy Bodies (DLB). LBDs are characterized by the presence of proteinaceous intracellular entities containing aggregates of  $\alpha$ -synuclein. The aggregation of these entities in specific brain regions leads to the formation of Lewy bodies and Lewy neuritis, usually accompanied by neurodegeneration in the affected areas (Beyer et al., 2009; Postuma et al., 2019).

$\alpha$ -synuclein is abundantly expressed throughout the human brain, with high levels in the neocortex, hippocampus, substantia nigra, thalamus and cerebellum (Kim et al., 2014). Although the exact function of  $\alpha$ -synuclein is unknown, it has been suggested that they may regulate the release of dopamine in controlling voluntary and involuntary movements or influence memory and cognitive function (Kim et al., 2014).

LBDs have a long prodromal interval, i.e. a period during which neurodegenerative symptoms are present, but full clinical disease has not yet developed (Berg et al., 2015). Identification of early stages in LBDs is crucial for the development of disease-modifying treatment, since the neurodegeneration may be possibly stopped or treated before the pathological cascades start.

Prodromal markers of LBDs are diverse and usually non-specific, with few exceptions such as idiopathic REM sleep behaviour disorder (Postuma et al., 2019). Nevertheless, recent research has focused on the state of mild cognitive impairment (MCI) in LBDs and it has been suggested that speech/voice disorders such as dysfluency, aperiodicity or irregular alternating motion rate can be identified in early stages of PD or DLB (Ash et al., 2012; Brabenec et al., 2017; Rusz et al., 2016).

Exploiting nowadays fashion of using voice assistants, which are implemented by almost all ICT companies in smartphones, computers and TVs, the use of voice and speech for research seems to have a promising future since access to databases of voice of millions of individuals, from all ages and genders, are expected to be available in the following years.

# Motivation

---

The motivation of this project is based on the fact that LBDs are a worldwide problem nowadays, affecting millions of individuals every year. Only in the U.S., it has been estimated that Lewy Body dementias affect more than 1.4 million individuals (National Institute of Aging, 2018) and the prevalence of Parkinson's disease was estimated to be 6.2 million people worldwide in 2015 (European Brain Council, n.d.; GBD 2015 Neurological Disorders Collaborator Group et al., 2017).

There is an essential necessity to improve current diagnosis methodology in order to identify prodromal stages in LBDs and new biomarkers and techniques must be explored so as to gain insight about the state of LBDs in a non-invasive, fast and affordable way.

Having worked with patients with tremor diseases during previous internships and with a special interest on Data Sciences, this project combines both fields and aims to offer a real-life solution for future research in diagnosis.

# Objectives

---

The main objective of this Final Bachelor Degree Project is to try to identify people at risk of developing LBDs based on an acoustic analysis of speech and voice.

The project aims to offer a real-life solution to a problem that affects millions of people worldwide and is expected to increase in prevalence in the following years.

In order to carry out this project in an organized, efficient way, individual objectives have been specified as follows:

1. Make a “state of the art” study to gain insight knowledge on LBDs, current methodology and techniques for diagnosis and recent research for diagnosis of LBDs.
2. Process and analyze acoustic data.
3. Explore different Machine Learning algorithms to identify acoustic features that discriminate between healthy controls and PD patients.
4. Propose a model, based on acoustic features, to identify people at risk of LBDs.

The long-term objective of this Project would be the incorporation of an acoustic test of speech and voice in current diagnostic criteria in hospitals and health-care points to improve present-day diagnosis outcomes.

# State of the art

---

## Lewy Bodies Diseases

As stated before, LBDs is a group of neurodegenerative diseases that include PD, PDD and DLB. These diseases, characterized by the aggregation of Lewy bodies and Lewy neurites in specific brain regions, differ in clinical features, symptomatology and diagnostic evaluation criteria. In this section, an in-depth research study will be presented in order to show the main characteristics of PD, PDD and DLB and their current clinical evaluation criteria.

## Parkinson's disease

Parkinson's disease is a multifactorial neurodegenerative disease characterized by the progressive impairment of voluntary motor controls.

The major pathological feature that correlates with PD is the progressive degeneration of dopaminergic neurons in the substantia nigra pars compacta, one of the nuclei constituting the basal ganglia. These neurons are involved in transmitting dopamine and their degeneration leads to dysfunction of the neuronal circuits that include motor cortical areas and the basal ganglia (Lotankar et al., 2017).

The disorder involves multiple motor and non-motor symptoms. The impairment of voluntary motor control might result in akinesia (failure of willed movement), bradykinesia (slowness of movement), hypokinesia (decreased bodily movement), postural instability, rigidity, stooped posture, and tremor at rest (Lotankar et al., 2017). These signs and symptoms are commonly present along with gait impairment, stiffness of the arms, legs, and trunk, poor balance and coordination, and bilateral vocal cord paralysis at the extreme and worsening level (Lotankar et al., 2017). These main motor symptoms, characteristic of PD, are collectively called "Parkinsonism".

Most common non-motor symptoms include Rapid Eye Movement (REM) Sleep Behavior Disorder (RBD), olfactory and autonomic dysfunction and mood disorders. These symptoms are often neglected, but might be of great benefit for early diagnosis as some of them, as in the case of RBD, have proven to be robust biomarkers of prodromal PD and other synucleinopathies (Le et al., 2017).

Currently, diagnosis of PD is based on clinical features from history and examination, and over time response to dopaminergic medication and development of motor symptoms (Rizek et al., 2016). Advancements in neuroimaging studies, such as Magnetic Resonance Imaging (MRI), are used to exclude other causes that might provoke Parkinsonism (Rizek et al., 2016). In addition, radionuclide imaging modalities like PET and SPECT, based on dopamine transporter ligands, have become the best approach to assess dopamine metabolism and deficiency, reaching up to 98% sensitivity and specificity to detect dopaminergic neuronal loss (Le et al., 2017). Other biomarkers for PD diagnosis include biochemical measurements of body fluids, such as  $\alpha$ -synuclein and DJ-1, and genetic biomarkers (Lotankar et al., 2017). These techniques are usually time-consuming and expensive, and thus clinical criteria almost always relies on clinical history and motor symptomatology. For these

reasons, there is an essential need to find new, fast, accurate and inexpensive biomarkers and methods that could improve and complement current diagnostic methodologies.

The Unified Parkinson's Disease Rating Scale (UPDRS) is the most commonly used rating tool to diagnose and evaluate PD progression. The UPDRS is composed of four main subscales, which includes (I) Mentation, Behavior and Mood, (II) Activities of Daily Living, (III) Motor Examination and (IV) Complications of Therapy. The effect of speech shows up only in two components by means of the patient's vocal output: in the second subscale for assessing whether it is apprehensible and in the third subscale for evaluating whether it is expressive during a conversation (Erdogdu Sakar et al., 2017).

The evaluation of speech in the UPDRS III is very general and non-specific. Hypokinetic Dysarthria (HD), a common speech disorder present in PD patients that will be further explained in this project, is a complex, multidimensional disorder that requires of a more specific, detailed analysis not covered in the current UPDRS III rating scale. In this project, an acoustic analysis will be carried out to perform a more accurate examination of speech and voice.

## Lewy Bodies dementias: Parkinson's disease with Dementia and Dementia with Lewy Bodies

Dementia with Lewy Bodies and Parkinson's disease dementia, also called Lewy Body dementias, are the second most common type of degenerative dementia in patients older than 65 years after Alzheimer's Disease (Gratwicke et al., 2015).

The main difference between PDD and LBD strikes in the relative timing of appearance of dementia and Parkinsonism. As such, PDD is considered a late complication of a well-established PD and has a cumulative prevalence of 75–90% of those with a disease duration of 10 years or more (Gratwicke et al., 2015). In contrast, DLB often occurs before or concurrently with parkinsonism or within 1 year of onset of motor symptoms (Walker et al., 2015).

The clinical features of DLB and PDD are similar and include hallucinations, cognitive fluctuations, and dementia in the setting of Parkinsonism. In addition, the cognitive domains that are impacted in DLB and PDD overlap substantially, with evident executive dysfunction and visual-spatial abnormalities, as well as variable impairment in memory capacities (Walker et al., 2015).

DLB and PDD exhibit multiple challenges for correct diagnosis. The biggest challenge in the diagnosis of DLBs is early diagnosis and differentiation from Alzheimer's disease (AD). In PDD, the main challenge is prediction and early identification of cognitive impairment in patients with established PD (Gomperts, 2016).

Imaging techniques such as CT or MRI scan are often part of the basic clinical examination in patients with suspected LBD or PDD. Regarding these techniques, the presence of occipital hypometabolism by means of FP-CIT SPECT has shown to be the most distinct finding in LBD (and PDD) compared with AD (Walker et al., 2015). Biological biomarkers, including plasma and cerebrospinal fluid  $\alpha$ -synuclein, have also shown promising results (Walker et al., 2015).

Diagnostic criteria for clinical diagnosis of PDD and DLB is based on clinical features and the timing of appearance of Parkinsonism. The central feature for diagnosis of possible or probable DLB is the

presence of dementia with progressive cognitive decline, accompanied by fluctuating cognition, recurrent visual hallucinations and/or spontaneous motor manifestations of Parkinsonism (Gomperts, 2016). Regarding PDD, the core features for diagnosis are a previous diagnosis of idiopathic Parkinson's disease and a dementia syndrome with insidious onset and slow progression, accompanied by one or more associated clinical features related with the cognitive and behavioral state (Gomperts, 2016).

## Current research in Lewy Bodies Diseases

Recent advances in the understanding and comprehension of PD and DLB are expected to impact future diagnosis and management of these diseases. These advances include the study of preclinical features, research in the concept of Mild Cognitive Impairment, and disease-modifying treatments (Gomperts, 2016).

Disease-modifying treatments of LBDs are more likely to succeed when started as early as possible; therefore, early diagnosis is a priority. Biomarkers such as constipation and loss of sense of smell might be present before cognitive and/or motor symptoms become evident, but they are non-specific. Only a few clinical features, like REM sleep behavior disorder and dopaminergic imaging abnormalities, have proved to be prodromal biomarkers of LBDs (Galasko, 2017).

Motor speech disorders have been long known to be present in patients with Parkinson's disease. Dysarthria is defined as "a group of related speech disorders that are due to disturbances in muscular control of the speech mechanism resulting from impairment of any of the basic motor processes involved in the execution of speech" (Godino-Llorente et al., 2017) and it is a common disorder in PD patients, with a prevalence up to 90% in the case of Hypokinetic Dysarthria (HD) (Godino-Llorente et al., 2017).

Hypokinetic dysarthria is mainly associated with dysfunctions of the basal ganglia control circuits, which are the loops that mediate between cognitive and motoric processes at the hippocampus (Godino-Llorente et al., 2017). More specifically, the basal ganglia regulates muscle tone, providing support for voluntary motor movements, controlling postural stability, and assisting in motor learning, among others (Godino-Llorente et al., 2017).

Phonation, articulation and prosody are the three main dimensions of speech affected in HD. Phonation refers to the physical process by which vocal folds produce sounds; articulation is the process by which sounds are created by the speech organs (lips, jaw, teeth, etc.); and prosody is concerned with properties of speech such as intonation, tone, stress and rhythm.

A variety of parameterization methods, using conventional and non-conventional speech features, have been developed for HD analysis. Thus far, conventional speech features have been used to describe the impairment of phonatory aspects of speech, speech quality deterioration, impairment of speech prosody, speech rate disturbances and impaired consonant articulation and tongue movement (Brabenec et al., 2017).

Non-conventional speech features, such as features based on Empirical Mode Decomposition, provide more precise HD identification; however, these are generally less clinically interpretable (Brabenec et al., 2017).

Brabenec et al., (2017), carried out a review of previous studies about HD in PD patients. Regarding the speech tasks used for early diagnosis of PD, most of the researchers used sustained phonation of vowels /a/, /e/, /i/, /o/, and /u/ (1 work analyzed the whole set of vowels, 1 work analyzed only vowel /i/ and another work analyzed a subset of vowels) and monologue and reading tasks. Furthermore, DDK evaluation and sentence repetition were also explored. Articulation and prosody dimensions of speech were found to be the most discriminative for early diagnosis of PD.

In addition to direct diagnosis of PD in its early stage, Rusz et al., (2016), published a study based on acoustic analysis of speech in a set of 16 patients with idiopathic RBD, who are generally at potential risk of developing Parkinson's disease, and 16 age- and sex- matched HC. They found that speech dysfunction was present in 88% of patients with RBD and articulatory impairment was the most prominent deficit.

## Theoretical Fundamentals

---

### Introduction to Machine Learning

Artificial Intelligence (AI) can be described as the attempt to develop human intelligence processes on computer systems. Machine Learning (ML) is an application of AI that provides computer systems the capacity to automatically learn from experience (data) without being specifically programmed. These algorithms are generally categorized as supervised or unsupervised.

Supervised ML algorithms require known input and output data in order to learn and predict future outputs. In this type of algorithms, the input data must belong to one or multiple classes so that the algorithm infers a mapping function for the prediction of future data. These algorithms are mainly used for classification and regression purposes.

In contrast, unsupervised ML algorithms are employed to find hidden patterns and/or structures from input data without the requirement of known labels or classes. Unsupervised algorithms are used for data clustering and association, among others.

In the framework of this project, different supervised ML algorithms will be employed. As stated before, this type of ML algorithms might be used for classification or regression aims. Regression models attempt to estimate numerical or continuous outputs from input data, while classification algorithms aim to infer discrete or categorical outputs. The former will be employed throughout the development of the project.

### Evaluation of the performance of ML algorithms

Evaluation of ML classification algorithms, and statistical classification tests in general, is a key step to assess the performance of the algorithms employed. The metrics assessed in this project, including classification accuracy, confusion matrix, Matthews correlation coefficient and area under ROC curve, will be further explained in this section.

Classification accuracy is defined as the ratio of correct predictions to the total number of predictions. This metric is very sensitive to the number of samples belonging to each class and, therefore, it is not useful when data is highly unbalanced.

A Confusion Matrix, as its name suggests, is a matrix that represents four different combinations between predicted and actual values within a dataset. Figure 1 illustrates a common Confusion Matrix.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 1. Structure of a Confusion Matrix

To further explain the basic concepts of the Confusion Matrix, a classification ML model with binary data i.e. data belongs to class 0, known as “negative”, or class 1, known as “positive”, will be assumed.

There are 4 important terms that can be inferred from a Confusion Matrix:

- True Positives (TP): It represents the number of correct positive predictions i.e. the number of samples whose value was 1 and the model has properly classified as 1.
- True Negatives (TN): It represents the number of correct negative predictions i.e. the number of samples whose value was 0 and the model properly classified as 0.
- False positive (FP): It represents the number of samples that were classified as positive (class 1) but were actually negative (class 0).
- False negative (FN): It represents the number of samples that were classified as negative (class 0) but were actually positive (class 1).

Accuracy can also be calculated from the Confusion matrix as described in Equation 1:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad \text{Equation 1}$$

The Sensitivity or True Positive Rate corresponds to the proportion of positive samples classified as positive (TP) with respect to all actual positive samples (TP+FN), explained in Equation 2. This metric can be interpreted as how well the model identifies people with the positive condition (1 or “True”).

$$Sensitivity = \frac{TP}{TP + FN} \quad \text{Equation 2}$$

On the other hand, Specificity or False Positive Rate is defined as the proportion of negative samples classified as positive (FP) with respect to all actual negative samples (TN+FP), displayed in Equation 3. This value is interpreted as the ability of the algorithm to properly identify healthy controls (0 or “False”).

$$Specificity = \frac{FP}{FP + TN} \quad \text{Equation 3}$$

The Matthews correlation coefficient (MCC) is used as a measure of the quality of an algorithm for binary classification. This metric is especially useful as it can be used even when data is highly unbalanced. The values of MCC are between -1 and +1; a coefficient of +1 represents perfect prediction, 0 no better than random and -1 indicates total disagreement between prediction and observations. The formulation of this coefficient is shown in Equation 4.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{(TP + FP)(TP + FN)(TN + FP)(TN + FN)} \quad \text{Equation 4}$$

The last metric evaluated in this project is the area under the ROC curve (AUC), one of the most widely used metrics for evaluation of ML models mainly employed in binary classification problems. The curve is represented by plotting the True Positive Rate vs. 1 - True Negative Rate (also called Recall) at different points between [0, 1].

The ROC curve can be interpreted as a probability curve and the AUC indicates the model overall performance in classifying samples among classes. The closer the AUC is to 1, the better the model performs, and an AUC of 0.5 represents a model which classifies no better than random.

## Data analysis and Machine Learning algorithms

In this section, the different methods and algorithms used throughout the project will be described. First, the concept of confidence intervals and their determination will be introduced. Later on, a brief introduction to linear regression will be made, followed by the description of the feature selection techniques and main classification algorithms employed in the project, which includes logistic regression and XGBoost.

### Confidence intervals

Diagnostic tests, as well as all modelling techniques, require of cutoff values for classifying cases as positive or negative. The determination of these values might be automatically performed by algorithms or manually calculated following a specific statistical method.

A simple and commonly used technique for calculating cutoff scores is the application of 95% confidence intervals (CI) of mean. In this method, mean and standard deviation (SD) of a known dataset are calculated to consequently determine an interval by subtracting and adding 2 x SD from and to the mean (also written as  $\mu \pm 2\sigma$ ) (Singh, 2006). Using this approach, the possibility of finding a value outside the interval is less than 5%.

Depending on the requirements of the modelling technique or test that is being performed, a different threshold might be established. Our project was made in collaboration with neurologists from the First Department of Neurology, Faculty of Medicine and St. Anne's University Hospital, Masaryk University, and their requirement was to obtain a model with 80% of specificity. Therefore, throughout this study, specificity will be prioritized over sensitivity i.e. the correct identification of healthy controls will be prioritized over the correct identification of patients since neurologists consider of greater importance the detection of patients who are not in risk of developing LBDs.

The dataset of positive cases (PD patients) will be considered for the calculation of the 95% CI in order not to lower specificity (in this way, sensitivity might be lowered due to the declaration of false positives cases). Considering the distribution of the PD and HC cohorts, the upper or lower limit of the interval ( $\text{mean} \pm 2\text{SD}$ ) will be established as the cutoff point depending on the specific feature distribution. This adaptation is required since distributions may vary from one feature to another.

## Feature selection approaches

In many pattern recognition applications, identifying the most characterizing features of the observed data is critical to minimize the classification error (Peng et al., 2005). Statistical analysis, generally classified as parametric or non-parametric, are widely used in research as they provide meaning to data and allow to draw conclusions from uncertain events, e.g. which subset of features characterizes a classification variable.

Parametric approaches can be more informational and interpretable, but should be used only under specific circumstances. On the other hand, non-parametric methods have the advantages that results are not affected by assumptions about the distribution nor by outliers (only signs and ranks are employed) and they can be used in small datasets, but might be less informative. In this project, the former will be employed to analyze the relevance of features.

### *Mann Whitney U test*

The Mann Whitney U test is a statistical, non-parametric method widely used for the comparison of two independent samples. In general terms, this test compares all samples from two independent populations to evaluate the null hypothesis, which assumes that the distribution of both populations is identical. It is usually compared with the t-test, but in this case the distribution does not have to be normal. In this project, due to the small dataset available, distribution assumptions will preferably not be made, thus the Mann Whitney U test will be used instead of the most common t-test.

Regarding this project, the fact that the distribution of two populations (PD and HC, for each feature) is different ( $p\text{-value} < 0.05$ ) is interpreted as that feature having high discriminant power for the classification of these populations.

### *Maximum Relevance Minimum Redundancy Feature Selection*

Most complex approaches that aim to identify a subset of features to characterize the target classification variable require the maximal statistical dependency (Max-Dependency) of the target class on the data distribution (Peng et al., 2005).

Due to the difficulties to compute Max-Dependency, multiple approaches are used, one of the most popular being the maximal relevance (Max-Relevance) feature selection. Relevance is usually characterized in terms of correlation or mutual information, of which the latter is used in the approach employed in this study.

Under the criteria of maximum-relevance minimum-redundancy, the selected features are required, individually, to have the largest mutual information (highest dependency) with the target class while being not redundant. The minimal redundancy condition is essential since features selected according to the max-relevance condition are usually mutually exclusive as the dependency between them is often large (Peng et al., 2005).

Thus, the result of this algorithm is a subspace of features, whose size can be defined, where features are sorted based on the Maximum Relevance - Minimum Redundancy criteria.

## Logistic Regression

Logistic Regression is a statistical model widely used in Machine Learning for binary classification problems. The model is based on the logistic function, also called sigmoid function, which is used to examine the relationship between a dichotomous (binary) dependent variable and one or more independent variables, which could be categorical or continuous.

The standard Logistic or Sigmoid function, described in Equation 5, is described within the real domain i.e. it can take any real input variable. The output or predicted value from a logistic function is always within the range  $[0, 1]$  and is interpreted as the probability that the given input sample belongs to a specific class.

$$\sigma = \frac{1}{1 + e^t} \quad \text{Equation 5}$$

Most often, a specific threshold is established for classification purposes so that all predicted outputs above the threshold are classified as 1 or “True” and outputs below the threshold are to be considered 0 or “False”. As stated before, in this project 80% specificity is desired and thus predicted outputs will be compared with the specific threshold that yields 80% specificity from the ROC curve.

## XGBoost

The last algorithm explored in this project is XGBoost. XGBoost or Extreme Gradient Boosting is a state-of-the-art algorithm based on tree boosting. More specifically, XGBoost implements a gradient boosting tree algorithm (Chen et al., 2016).

The mathematical comprehension of XGBoost goes beyond the scope of this project, but the main concepts of the algorithm will be presented in this section.

First, a short discussion about Decision Tree algorithms will be made, followed by a brief explanation of the concept of bagging and random forests, and a final introduction to Gradient Boosting and its relation with XGBoost.

Decision tree is a type of supervised Machine Learning algorithm used for classification and regression problems. A Decision Tree algorithm might be interpreted as a tree where each node represents a feature, each link (branch) represents a decision and each leaf represents an outcome, that can be categorical or continuous (Sanjeevi, 2017).

The main idea behind Decision Trees can be described in three steps (Navlani, 2018):

- First, the most discriminative attribute or feature is selected based on a specific metric, such as Gini Index or Information Gain.
- Then, a Decision Node is created with the selected feature to split the data.
- In the third step, the process is recursively performed until a stop condition from the criteria occurs, such as the overpass of the maximum number of permitted iterations or no features availability.

The flowchart-like structure of Decision Trees mimics human-thinking processes, what makes them easy to understand and interpret.

Ensemble learning is based on the idea that combination of multiple weak Decision Trees classifiers can create a strong new classifier, while also reducing variance and increasing the robustness of the model. There are two main ensemble methods: bagging and boosting.

Bagging or Bootstrap Aggregation is a technique that combines predictions from multiple, parallel classifiers through a majority voting mechanism. This idea is the basis of Random Forest algorithms, where a subset of features are randomly chosen to build a collection of Decision Trees known as Forest, hence the name of the algorithm (Rocca, 2019).

Boosting is another ensemble technique for improving model predictions by sequentially training models that correct the errors of previous models. Models are sequentially added until no further improvements can be made (Dabbura, 2017).

Gradient Boosting is a special case of boosting where errors are minimized from one model to another using a Gradient Descent (GD) algorithm. In a GD algorithm, the parameters of the model are updated at each iteration following the opposite direction of the gradient of the function at a specific learning rate (Dabbura, 2017). GD is a first-order optimization algorithm since it only considers the first derivative of the function.

XGBoost, as stated at the beginning of this section, is an implementation of a Gradient Boosted Decision Tree especially focused on computational speed and performance. The improvements and advancements of this algorithm rely on its parallelized implementation and hardware optimization by means of distributed computing, out-of-core computing and cache optimization (Chen et al., 2016).

# Methodology

## Participants

In the framework of this project, 36 Czech native PD patients with age of (mean  $\pm$  std)  $76.64 \pm 16.51$  years (11 females  $79.09 \pm 13.57$  years, 25 males  $77.15 \pm 17.29$  years), 29 healthy controls HC with age of  $68.56 \pm 6.11$  years (20 females  $69.00 \pm 6.32$  years, 9 males  $67.6 \pm 5.50$  years), and 37 subjects who were at risk of developing LBDs with age of  $65.19 \pm 4.97$  years (24 females  $66.09 \pm 4.67$  years, 13 males  $63.62 \pm 5.09$  years) were enrolled at the First Department of Neurology, St. Anne's University Hospital in Brno, Czech Republic. For demographic and clinical data of PD patients, see Table 1.

**Table 1.** Clinical characteristics of PD patients

	Mean	Std	Min	1Q	Median	3Q	Max
PD dur (y.)	5.56	5.08	0.5	1.5	4	8.5	20
UPDRS III	14.47	6.38	6	9.75	14	17.5	29
LED (mg/day)	854.24	482.53	80	500	750	1251.75	1850
Age (y)	76.64	16.51	55	66.5	73	80	114

UPDRS III – Unified Parkinson's disease rating scale, part III (motor examination), LED – Levodopa equivalent dose.

The participants at risk of LBDs were identified based on a screening questionnaire containing several risk factors, e.g. the REM sleep behaviour disorder. None of the PD patients had a disease affecting the central nervous system other than PD. These patients were examined on their regular dopaminergic medication approximately 1 hour after the L-dopa dose. All participants signed an informed consent form that has been approved by the local ethics committee.

## Acoustic analysis

Speech/voice of the enrolled participants was recorded by a large capsule cardioid microphone M-AUDIO Nova and sampled at  $f_s = 16$  kHz.

All participants followed the same protocol, which included the speech tasks described in Table 2. The speech/voice disorders quantified by acoustic features are explained in Table 3.

Clinicians were asked to explain carefully every task to the patient before performing the test and to not interrupt them during the performance. The acquisition of the recording was done in a quiet room with low echo and with no background noise such as sound of street, air condition, computer fan, clock, telephone, medical devices, etc. In addition, the text of TSK2 was printed on a card/board so as it would not make any noise when being held.

**Table 2.** Speech Tasks.

<b>Label</b>	<b>Speech task</b>	<b>Description</b>
TSK1	Monologue	Monolog, at least 90 s long without interruption of a clinician. The participants will be instructed to speak about their hobbies, family, job, actual date activity, etc.
TSK2	Reading	Reading a short text. The patient can read the text for her-/himself in advance.
TSK3	Sustained phonation	Approximately 3-s (not longer than 5 s) sustained vowel of /a/ at a comfortable pitch and loudness. Performed on one breath.
TSK4	Sustained phonation	Approximately 3-s (not longer than 5 s) sustained vowel of /i/ at a comfortable pitch and loudness. Performed on one breath.
TSK5	Sustained phonation	Approximately 3-s (not longer than 5 s) sustained vowel of /u/ at a comfortable pitch and loudness. Performed on one breath.
TSK6	Sustained phonation	Sustained phonation of /a/ at a comfortable pitch and loudness as constant and long as possible, at least 5 s. Performed on one breath.
TSK7	Diadochokinetic task	Rapid steady /pa/-/ta/-/ka/ syllables repetition as constant and long as possible, repeated at least 5 times. Performed on one breath.

**Table 3.** Feature Set

HD/AOS dimension and specific disorder	Speech tasks	Acoustic feature	Feature definition
<b>Phonation</b>			
Airflow insufficiency	Sustained phonation (TSK6)	MPT	Maximum phonation time, aerodynamic efficiency of the vocal tract measured as the maximum duration of the prolonged vowel.
Irregular pitch fluctuations	Sustained phonation (TSK3–6)	relF0SD	Standard deviation of fundamental frequency relative to its mean, variation in frequency of vocal fold vibration.
Microperturbations in frequency	Sustained phonation (TSK3–6)	Jitter (PPQ)	Frequency perturbation, extent of variation of the voice range. Jitter is defined as the variability of the F0 of speech from one cycle to the next.
Microperturbations in amplitude	Sustained phonation (TSK3–6)	Shimmer (APQ)	Amplitude perturbation, representing rough speech. Shimmer is defined as the sequence of maximum extent of the signal amplitude within each vocal cycle.
Increased noise	Sustained phonation (TSK3–6)	HNR	Harmonics-to-noise ratio, the amount of noise in the speech signal, mainly due to incomplete vocal fold closure. HNR is defined as the amplitude of noise relative to tonal components in speech.
Aperiodicity	Sustained phonation (TSK3–6)	DUV	Degree of unvoiced segments, the fraction of pitch frames marked as unvoiced.
Tremor of jaw	Sustained phonation (TSK3–6)	relF1SD, relF2SD	Standard deviation of first (F1) and second (F2) formant relative to its mean. Formants are related to resonances of the oro-naso-pharyngeal tract and are modified by position of tongue and jaw.
<b>Articulation</b>			
Decreased tongue movement (imprecise vowels)	Monologue, reading (TSK1–5)	VAI	Vowel articulation index, based on formant centralization, defined as $VAI = (F1a + F2i)/(F1i + F1u + F2a + F2u)$ .

Rigidity of tongue and jaw	Monologue, reading (TSK1–2)	relF1SD, relF2SD	Standard deviation of first (F1) and second (F2) formant relative to its mean.
Slow alternating motion rate	Diadochokinetic task (TSK7)	DDK rate	Diadochokinetic rate, representing the number of syllable vocalizations per second.
Irregular alternating motion rate	Diadochokinetic task (TSK7)	DDK reg	Diadochokinetic regularity, defined as the standard deviation of distances between following syllables nuclei.
<b>Prosody</b>			
Monoloudness	Monologue, reading (TSK1–2)	relSEOSD	Speech loudness variation, defined as a standard deviation of intensity contour relative to its mean after removing silences exceeding 50 ms.
Monopitch	Monologue, reading (TSK1–2)	relF0SD	Pitch variation, defined as a standard deviation of F0 contour relative to its mean.
Inappropriate silences	Reading (TSK2)	SPIR	Number of pauses relative to total speech time after removing periods of silence lasting less than 50 ms.
Higher proportion of silence time	Reading (TSK2)	PPR	Percentual pause ratio, defined as total duration of silences (longer than 50 ms)/total duration of speech.
Longer duration of silences	Reading (TSK2)	DurMED	Median duration of silences longer than 50 ms.
Higher variability of silence duration	Reading (TSK2)	DurMAD	Median absolute deviation of silence duration (silences longer than 50 ms).
Unnatural speech rate	Reading (TSK2)	AR	Number of speech sounds produced per second after pauses longer than 50 ms were removed.

## Statistical Analysis

### Univariate analysis

Mean and standard deviation values were calculated for each individual feature using the complete set of participants (PD + HC) and for the PD and HC cohorts separately to analyze data distribution. In addition, boxplots were employed to visualize distribution per feature in the PD and HC cohorts.

In the first approach, in order to identify discriminative features with at least 80% of specificity, cut-off scores were determined based on the population distribution of particular features. On this basis, cut-off scores were calculated by means of 95% confidence intervals.

In a second stage, we fitted individual features in univariate logistic regression models in order to evaluate their discriminant power to differentiate between HC and PD subjects. The performance of these models was assessed according to their MCC, AUC, sensitivity and specificity.

To complete the univariate analysis, two approaches were implemented in order to evaluate the importance or relevance of each feature. The first approach consisted of a filtering feature selection algorithm that was employed to sort features based on the maximum relevance, minimum redundancy criteria. The algorithm required of a preprocessing step since data had to be previously discretized. The second approach comprised a Mann-Whitney U test, which was performed in the whole dataset to identify potential discriminant features ( $p\text{-value} < 0.05$ ).

### Multivariate analysis

#### *Multivariate logistic regression model*

In accordance with the results from the univariate analysis and feature selection algorithms, features identified as most discriminant were selected. Different multivariate regression models were fitted with all possible three-feature combinations among all features selected to identify the optimal combinations of features. As previously, these models were evaluated according to their MCC, AUC, sensitivity, and specificity.

#### *XGBoost*

In the last section, a second ML algorithm approach was employed for data classification. An XGBoost model was explored for this purpose, with the aim of compare and improve our previous results.

First, a random search with a cross-validation approach was used to find the best hyperparameters of the XGBoost classifier for our dataset. The hyperparameters considered included the following parameters and distributions: learning rate: [0.001, 0.01, 0.1, 0.2, 0.3]; gamma: [0, 0.25, 0.50, 1.0]; max\_depth: [4, 6, 8, 10, 20]; subsample: [0.5, 0.6, 0.7, 0.8, 0.9, 1.0], colsample\_bylevel and colsample\_bytree: [0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]; and min\_child\_weight: [0.5, 1.0, 3.0, 5.0, 7.0, 10.0].

In addition, multiple scoring methods (roc\_auc, F1 score and balanced accuracy) were explored, and 100 number of iterations and all processors were employed.

A second implementation using XGBoost was performed, in which an XGBoost classifier with the best hyperparameters found for our data was fitted with all possible combinations of our data avoiding 0, 1, and 2 samples. This procedure was performed in order to dismiss possible outliers that could disrupt the performance of the algorithm. The final model was employed for the identification of subjects at high risk of LBDs.

## Results

---

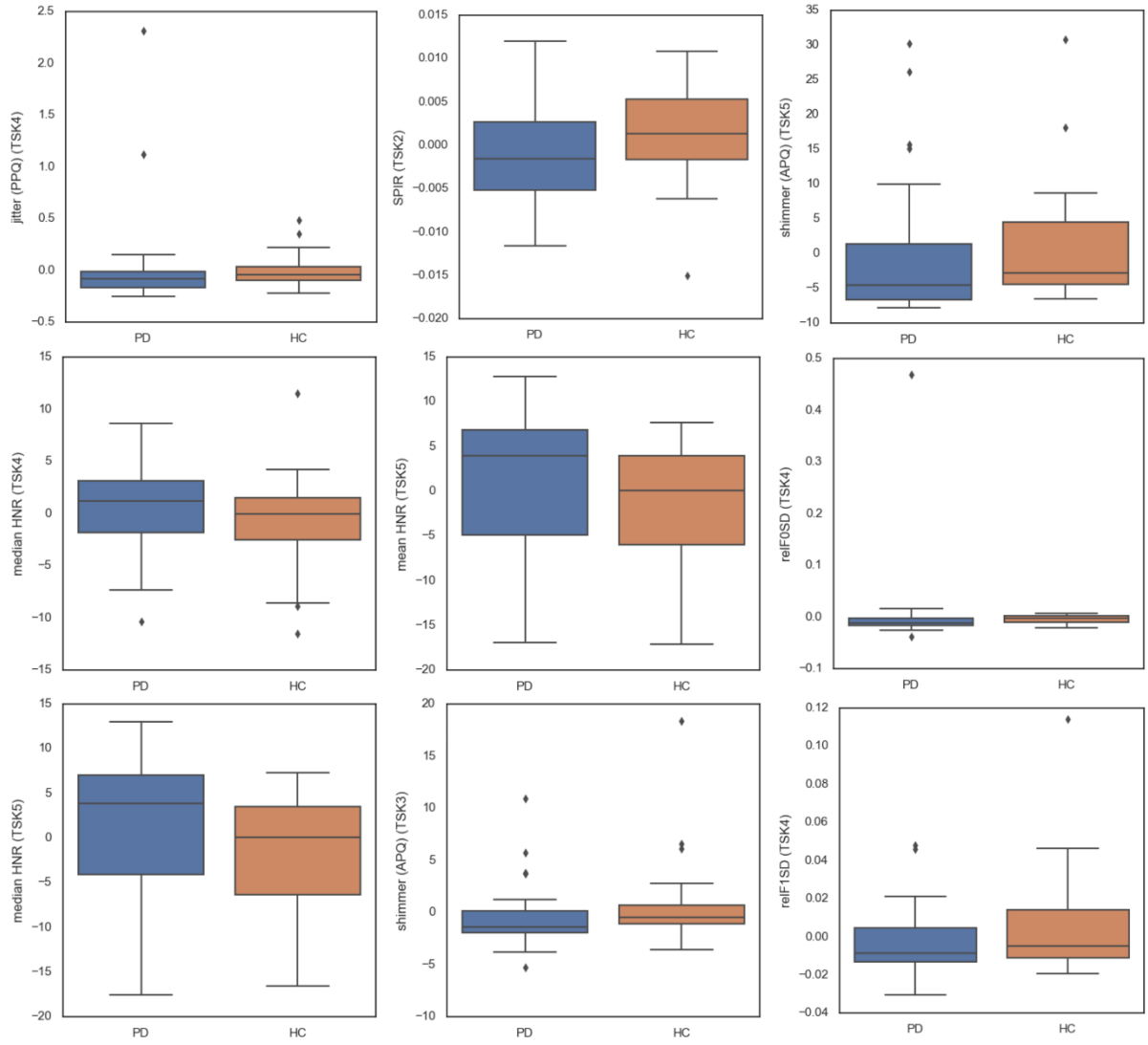
Results of the univariate regression analysis are summarized in Table 4. No feature was found to provide 80% specificity by means of 95% CI. However, according to the p-value obtained from the Mann-Whitney U test, relF0SD (TSK4), shimmer (APQ) (TSK5), shimmer (APQ) (TSK3), median HNR (TSK5), mean HNR (TSK5), SPIR (TSK2), jitter (PPQ) (TSK4), median HNR (TSK4) and relF1SD (TSK4) have significant distributional differences ( $p\text{-value} < 0.05$ ) and, thereafter, might be considered as potentially discriminative features. In addition, phonation was found to be the speech dimension with most discriminant power.

Regarding distributional differences, the features jitter (PPQ) (TSK4), SPIR (TSK2), shimmer (APQ) (TSK3), shimmer (APQ) (TSK5), relF0SD (TSK4) and relF1SD (TSK4) exhibited lower median values in the PD cohort than HC. On the other hand, PD subjects generally had higher median values in median HNR (TSK4), mean HNR (TSK5) and median HNR (TSK5) with respect to HC (see Figure 2).

**Table 4.** Results of the univariate regression analysis.

Task	Speech Dimension	Acoustic Feature	Mean + STD (PD)	Mean + STD (HC)	AUC [%]	MCC	Sensitivity [%]	Specificity [%]	Accuracy [%]	MRMR Importance	p-value
TSK4	Phonation	relF0SD	0.00 ± 0.08	0.00 ± 0.01	41.76	0.00	100.00	0.00	55.38	16	0.0096
TSK5	Phonation	shimmer (APQ)	-0.79 ± 9.26	0.98 ± 7.93	62.26	-0.01	88.89	10.34	53.85	21	0.0096
TSK3	Phonation	shimmer (APQ)	-0.53 ± 2.86	0.66 ± 3.95	64.75	0.05	86.11	17.24	55.38	49	0.0182
TSK5	Phonation	median HNR	1.33 ± 7.72	-1.65 ± 6.32	64.27	0.21	77.78	41.38	61.54	7	0.0243
TSK5	Phonation	mean HNR	1.25 ± 7.61	-1.55 ± 6.39	62.74	0.21	77.78	0.00	61.54	39	0.0338
TSK2	Articulation	SPIR	0.00 ± 0.01	0.00 ± 0.01	53.45	0.00	100.00	0.00	55.38	41	0.0348
TSK4	Phonation	jitter (PPQ)	0.01 ± 0.45	-0.01 ± 0.16	47.22	-0.11	97.22	24.14	53.85	26	0.0402
TSK4	Phonation	median HNR	0.68 ± 4.05	-0.84 ± 4.49	59.58	-0.01	75.00	0.00	52.31	12	0.0437
TSK4	Phonation	relF1SD	0.00 ± 0.02	0.01 ± 0.03	56.13	0.00	100.00	0.00	55.38	2	0.0488
TSK6	Phonation	RelF2SD	-0.02 ± 0.11	0.02 ± 0.13	54.79	0.00	100.00	0.00	55.38	1	0.0544

STD – standard deviation, HC – healthy controls, PD – Parkinson’s disease patients, AUC – area under curve, MCC – Matthews correlation coefficient, relF0SD – relative standard of the fundamental frequency, shimmer (APQ) – amplitude perturbation, HNR – harmonic-to-noise ratio, SPIR – speech index of rhythmicity, jitter (PPQ) – frequency perturbation, relF1SD and relF2SD – relative standard of the 1<sup>st</sup> and 2<sup>nd</sup> formant.



**Figure 2.** Boxplots of the PD and HC cohorts per feature.

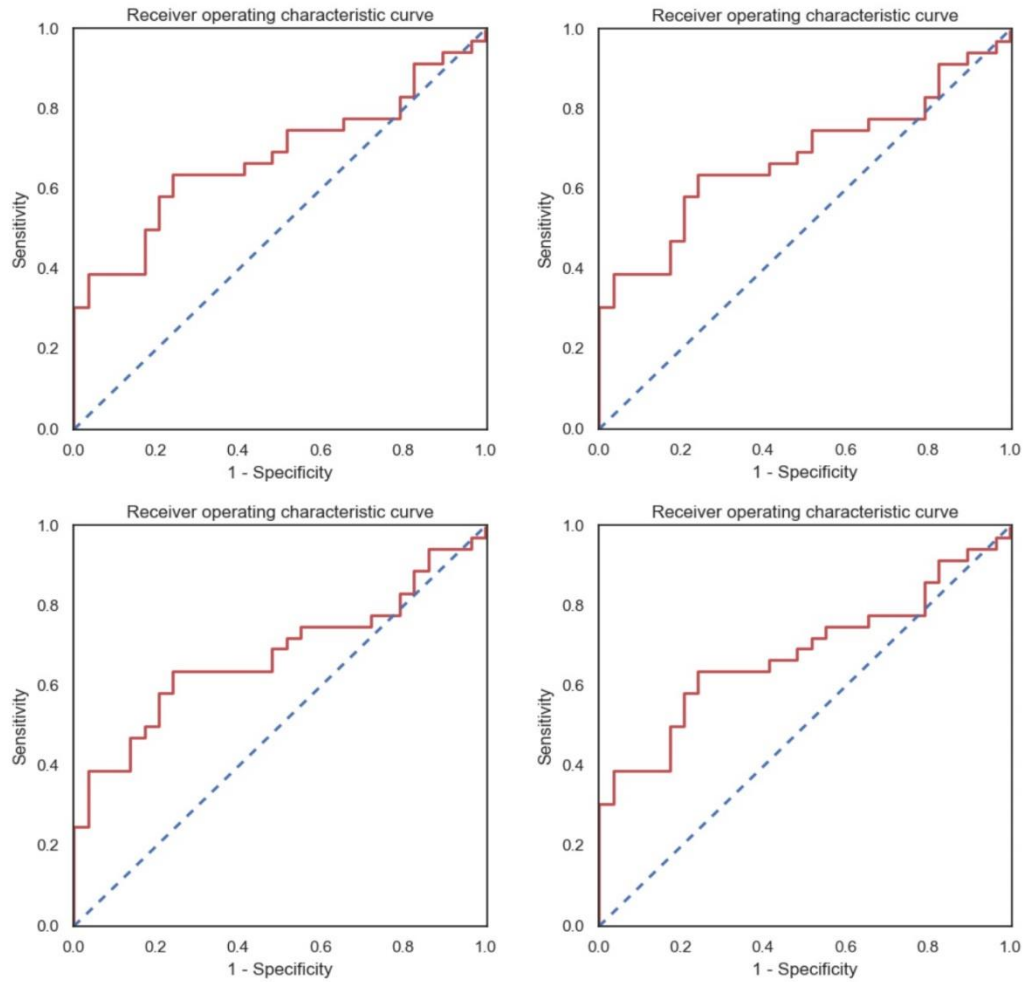
With respect to the multivariate regression analysis, multiple logistic regression models provided adequate results satisfying the initial requirements; more specifically, models combining the features shimmer (APQ) (TSK3), median HNR (TSK5) and either mean HNR (TSK5), relF0SD (TSK4), relF1SD (TSK4), SPIR (TSK2) or jitter (PPQ) (TSK4) (see Table 5).

For the purposes of this project, the models combining shimmer (APQ) (TSK3), median HNR (TSK5) and either relF0SD (TSK4), relF1SD (TSK4), SPIR (TSK2) or jitter (PPQ) (TSK4) were found to be optimal, achieving 79.32% specificity and 58.33% sensitivity (see ROC curves in Figure 3).

**Table 5.** Results of the multivariate analysis

<b>Model</b>	<b>AUC [%]</b>	<b>MCC</b>	<b>Sensitivity [%]</b>	<b>Specificity [%]</b>	<b>Accuracy [%]</b>
['shimmer (APQ) (TSK3)', 'median HNR (TSK5)', 'mean HNR (TSK5)']	70.11	0.17	75.00	41.38	53.85
['relF0SD (TSK4)', 'shimmer (APQ) (TSK3)', 'median HNR (TSK5)']	68.30	0.10	75.00	34.48	53.85
['shimmer (APQ) (TSK3)', 'median HNR (TSK5)', 'relF1SD (TSK4)']	68.30	0.10	75.00	34.48	53.85
['shimmer (APQ) (TSK3)', 'median HNR (TSK5)', 'SPIR (TSK2)']	68.20	0.10	75.00	34.48	53.85
['shimmer (APQ) (TSK3)', 'median HNR (TSK5)', 'jitter (PPQ) (TSK4)']	67.91	0.07	75.00	31.03	53.85
['relF0SD (TSK4)', 'shimmer (APQ) (TSK3)', 'mean HNR (TSK5)']	67.82	0.17	77.78	37.93	53.85
['shimmer (APQ) (TSK3)', 'mean HNR (TSK5)', 'SPIR (TSK2)']	67.72	0.17	77.78	37.93	53.85
['shimmer (APQ) (TSK3)', 'mean HNR (TSK5)', 'relF1SD (TSK4)']	67.72	0.17	77.78	37.93	53.85
['shimmer (APQ) (TSK3)', 'mean HNR (TSK5)', 'jitter (PPQ) (TSK4)']	67.05	0.10	77.78	31.03	53.85
['median HNR (TSK5)', 'mean HNR (TSK5)', 'relF1SD (TSK4)']	66.67	0.14	75.00	37.93	53.85

AUC – area under curve, MCC – Matthews’s correlation coefficient, relF0SD – relative standard of the fundamental frequency, shimmer (APQ) – amplitude perturbation, HNR – harmonic-to-noise ratio, SPIR – speech index of rhythmicity, jitter (PPQ) – frequency perturbation, relF1SD – relative standard of the 1<sup>st</sup> formant.



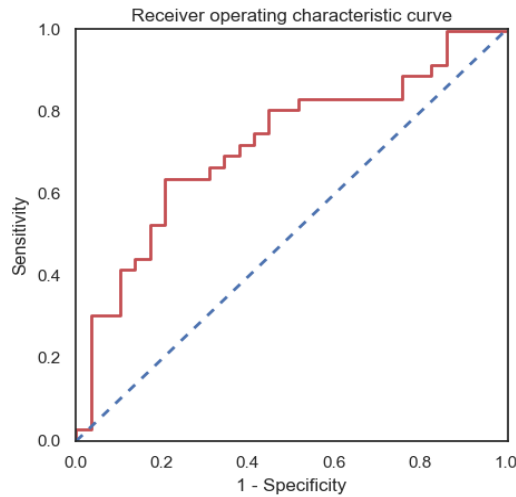
**Figure 3.** ROC curves from the multivariate analysis. Models consisting of shimmer (APQ) (TSK3), median HNR (TSK5) and: top-left, relF1SD (TSK4) and relF1SD (TSK4); top-right, SPIR (TSK2); bottom-left, jitter (PPQ) (TSK4); and bottom-right, relFOSD (TSK4).

The ROC curve of the last model employed, using an XGBoost classifier, is displayed in Figure 4. The hyperparameters found to fit better our data include a scorer based on the roc curve; learning rate: 0.3; gamma: 0; maximum depth: 20; minimum child weight: 3; colsample by level: 0.7; and colsample by tree: 0.5. The average values that described the performance of this model are shown in Table 6.

**Table 6.** Results of the XGBoost Classifier

Model	AUC [%]	MCC	Sensitivity [%]	Specificity [%]	Accuracy [%]
XGBoost	69.70	0.32	67.00	66.00	66.00

AUC – area under curve, MCC – Matthews correlation coefficient



**Figure 4.** ROC curve of the XGBoost classifier.

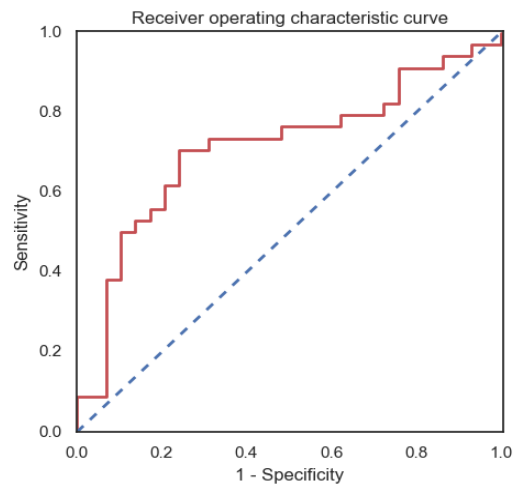
Lastly, a process denominated “surgery” was applied to our data in order to dismiss potential samples that could disrupt the model performance (see outliers in Figure 2). Three samples were dismissed, all of them belonging to the PD cohort. Figure 5 displays the new ROC curve obtained and Table 7 summarizes the metrics of the performance of the model.

**Table 7.** Results of the XGBoost classifier with surgery

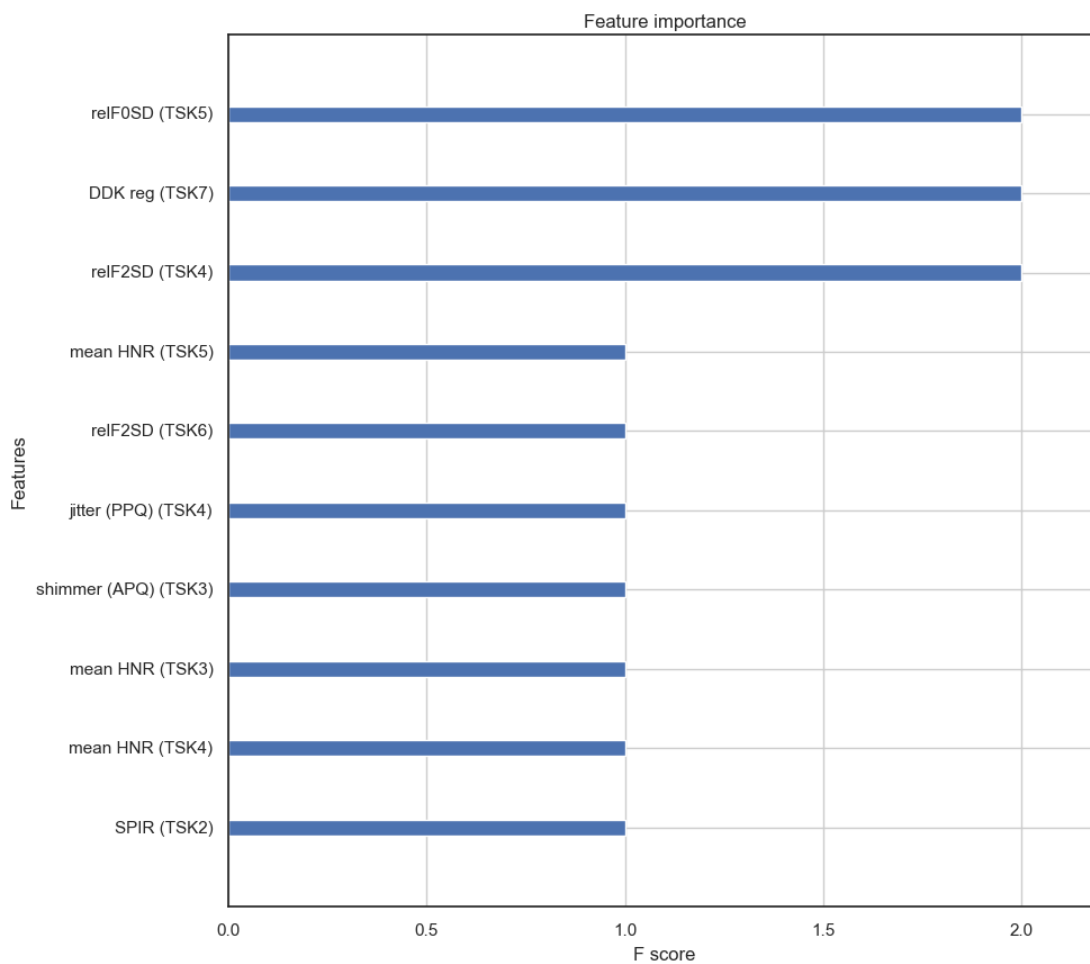
Model	AUC [%]	MCC	Sensitivity [%]	Specificity [%]	Accuracy [%]
XGBoost & surgery	69.70	0.46	71.00	76.00	73.00

AUC – area under curve, MCC – Matthews’s correlation coefficient

The first ten features found to be the most discriminative in the XGBoost (with surgery) classifier are shown in Figure 6. Using this former model, which provide the best trade-off between specificity and sensitivity (with 76% specificity and 71% sensitivity, see Table 7), only thirteen participants were confirmed to be at risk of LBDs.



**Figure 5.** ROC curve of the XGBoost classifier with surgery.



**Figure 6.** Top most discriminant features of the XGBoost (with surgery) classifier.

## Discussion

---

The most discriminative features, according to the results from the univariate analysis, are based on the fundamental frequency extracted from the sustained phonation of vocal /i/, the microperturbations in amplitude within each vocal cycle extracted from the sustained phonation of vocals /a/ and /u/, and the harmonics-to-noise ratio measuring increased noise extracted from the sustained phonation of vocal /i/ and /u/.

Generally, the fundamental frequency is related to the variation of frequency in vocal fold vibration, what results in abnormal increases in pitch voice. Irregular pitch fluctuations have been previously identified in Parkinson's disease patients as a result of the stiffness in the vocal folds (Zhang et al., 2005) and the fundamental frequency has been proved to have high discriminant power during the sustained phonation of vocal /i/ by other studies (Bocklet et al., 2011). In our findings, this feature was found to be discriminant, but the HC cohort showed slightly higher median values than PD patients, contrary to expected.

Microperturbations in amplitude was also found to be a discriminant feature between PD and HC cohorts during the sustained phonation of vocal s/a/ and /u/, finding supported by other studies as well (Rusz et al., 2011).

Features related with the Harmonics-to-Noise (HNR) ratio, which measures increased noise due to the incomplete closure of vocal folds, were also identified to be discriminant during the sustained phonation of vocals /i/ and /u/. These types of features are widely used for their high discriminant power and clinical interpretability (Orozco-Arroyave et al., 2016; Rusz et al., 2011; Sakar et al., 2013). The median values within the PD cohort were generally higher than the HC, contrary to what was expected.

In addition, contrarily to literature, the monologue and reading tasks did not provide relevant information for the identification of PD patients and healthy controls. Indeed, the speech index of rhythmicity (SPIR) extracted from the reading task was the only feature found to be discriminant.

The important role of SPIR in the acoustic analysis of hypokinetic dysarthria was identified by Rektorova et al., who used this feature to predict mild cognitive impairment or dementia in PD patients (Rektorova et al., 2016). In view of the study, it can be assumed that this prosodic feature quantifying inappropriate silences is somehow associated with cognitive decline in the PD patients. The results from the data distribution among the PD and HC cohorts showed lower median values in the PD cohort with respect to the HC individuals, as expected according to existing literature.

The optimal combination of features using a multivariate logistic regression approach consisted of shimmer (APQ) (TSK3), median HNR (TSK5) and either relF0SD (TSK4), relF1SD (TSK4), jitter (PPQ) (TSK4) or SPIR (TSK2), achieving 79.00 % specificity and 58.00 % sensitivity.

The XGBoost classifier, after the hyper-parameterization fitting, improved the performance of previous models achieving 79% specificity and 64% sensitivity. Due to outliers in our data, a post-processing step to dismiss samples that could drastically disrupt the model performance was carried out. Three samples from the PD cohort were dismissed in this procedure.

The new XGBoost classifier improve its performance by reaching the best trade-off between sensitivity and specificity (76% specificity and 71% sensitivity); MCC: 0.46, accuracy: 73%. Based on this former model, 13 subjects were confirmed to be at risk of developing LBDs.

In accordance with the results from the univariate analysis, features measuring microperturbations in amplitude (shimmer (APQ) (TSK3)) and frequency (jitter (PPQ) (TSK4)), increased noise (mean HNR (TSK5)), inappropriate silences (SPIR (TSK2)), and tremor of jaw (relF2SD (TSK6)) were identified to be among the most discriminant features for the identification of PD patients and HC subjects.

As it was expected, the quantification of irregular alternating motion rate during the diadochokinetic task (DDK reg (TSK7)) was identified as a relevant, discriminant measure for classification purposes. Diadochokinetic regularity can be defined as the standard deviation of distances between following syllables nuclei and is related with the capacity to do the correct occlusion of the oral cavity.

Recent studies have shown promising results proving DDK evaluation as a robust, accurate method for discriminating between PD and HC speakers, achieving AUC values above 95.00 % in multi-language studies (Orozco-Arroyave et al., 2016).

Both the univariate analysis and the XGBoost found the most discriminative features measuring phonation through the sustained phonation of vocal /a/, /i/, and /u/ (TSK3-6). Only SPIR (TSK2), which is directly related to articulation, was found to be discriminant, accompanied by DDK reg (TSK7) in the analysis performed by the XGBoost algorithm.

## Conclusion

---

In this study, we identified acoustic features that discriminate between PD patients and HC and used these parameters to train different models (multivariate logistic regression and XGBoost models) that identified subjects at high risk of developing LBDs (achieving 76 % specificity and 71 % sensitivity). These subjects, in comparison to HC, were found to be associated mainly with phonetic and articulatory disorders.

The accuracy of the identification of subjects at high risk of developing LBDs will be evaluated in future, as the thirteen participants identified will be longitudinally monitored and clinically examined.

This work has several limitations, including a small cohort of participants and different severity of PD patients (the severity of PD patients was not included during the development of the different models). Therefore, we cannot generalize the results, but rather consider them as pilot ones.

On the other hand, to the best of our knowledge, it is the first work dealing with the identification of prodromal state of LBDs based on the acoustic analysis of speech/voice, and we hope that our findings will help in further research in this field of science.

## Project acknowledgments

---

This research was funded by the ERASMUS+ program, European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 734718 (CoBeN), the grant of the Czech Ministry of Health 16-30805A (Effects of non-invasive brain stimulation on hypokinetic dysarthria, micrographia, and brain plasticity in patients with Parkinson's disease) and grant LO1401. For the research, infrastructure of the SIX center was used.

The project was carried out during a mobility exchange within the framework of an ERASMUS+ traineeship at the Brain Disease Analysis Laboratory, Department of Telecommunications, Brno University of Technology (Czech Republic), in collaboration with the First Department of Neurology, Faculty of Medicine and St. Anne's University Hospital (Masaryk University) and Applied Neuroscience Research Group, Central European Institute of Technology–CEITEC (Masaryk University).

## References

---

- Ash, S., McMillan, C., Gross, R. G., Cook, P., Gunawardena, D., Morgan, B., ... Grossman, M. (2012). Impairments of speech fluency in Lewy body spectrum disorder. *Brain and Language*, 120(3), 290–302. <https://doi.org/10.1016/j.bandl.2011.09.004>
- Berg, D., Postuma, R. B., Adler, C. H., Bloem, B. R., Chan, P., Dubois, B., ... Deuschl, G. (2015). MDS research criteria for prodromal Parkinson's disease. *Movement Disorders*, Vol. 30, pp. 1600–1611. <https://doi.org/10.1002/mds.26431>
- Beyer, K., Domingo-Sàbat, M., & Ariza, A. (2009). Molecular pathology of lewy body diseases. *International Journal of Molecular Sciences*, Vol. 10, pp. 724–745. <https://doi.org/10.3390/ijms10030724>
- Bocklet, T., Nöth, E., Stemmer, G., Ruzickova, H., & Rusz, J. (2011). Detection of persons with Parkinson's disease by acoustic, vocal, and prosodic analysis. *2011 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2011, Proceedings*, 478–483. <https://doi.org/10.1109/ASRU.2011.6163978>
- Brabenec, L., Mekyska, J., Galaz, Z., & Rektorova, I. (2017). Speech disorders in Parkinson's disease: early diagnostics and effects of medication and brain stimulation. *Journal of Neural Transmission*, Vol. 124, pp. 303–334. <https://doi.org/10.1007/s00702-017-1676-0>
- Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. <https://doi.org/10.1145/2939672.2939785>
- Dabbura, I. (2017). Gradient Descent Algorithm and Its Variants – Towards Data Science. Retrieved April 15, 2019, from <https://towardsdatascience.com/gradient-descent-algorithm-and-its-variants-10f652806a3>
- Erdogdu Sakar, B., Serbes, G., & Sakar, C. O. (2017). Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease. *PloS One*, 12(8), e0182428. <https://doi.org/10.1371/journal.pone.0182428>

- European Brain Council. (n.d.). Parkinson's disease | EBC. Retrieved May 28, 2019, from <https://www.braincouncil.eu/library/disease-fact-sheets/parkinsons-disease/>
- Galasko, D. (2017). Lewy Body Disorders. *Neurologic Clinics*, Vol. 35, pp. 325–338. <https://doi.org/10.1016/j.ncl.2017.01.004>
- GBD 2015 Neurological Disorders Collaborator Group, V. L., Abajobir, A. A., Abate, K. H., Abd-Allah, F., Abdulle, A. M., Abera, S. F., ... Vos, T. (2017). Global, regional, and national burden of neurological disorders during 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet. Neurology*, 16(11), 877–897. [https://doi.org/10.1016/S1474-4422\(17\)30299-5](https://doi.org/10.1016/S1474-4422(17)30299-5)
- Godino-Llorente, J. I., Shattuck-Hufnagel, S., Choi, J. Y., Moro-Velázquez, L., & Gómez-García, J. A. (2017). Towards the identification of Idiopathic Parkinson's Disease from the speech. New articulatory kinetic biomarkers. *PLoS ONE*, 12(12). <https://doi.org/10.1371/journal.pone.0189583>
- Gomperts, S. N. (2016). Lewy body dementias: Dementia with lewy bodies and Parkinson disease dementia. *CONTINUUM Lifelong Learning in Neurology*, Vol. 22, pp. 435–463. <https://doi.org/10.1212/CON.0000000000000309>
- Gratwicke, J., Jahanshahi, M., & Foltynie, T. (2015). Parkinson's disease dementia: A neural networks perspective. *Brain*, Vol. 138, pp. 1454–1476. <https://doi.org/10.1093/brain/awv104>
- Kim, W. S., Kagedal, K., & Halliday, G. M. (2014). Alpha-synuclein biology in Lewy body diseases. *Alzheimer's Research and Therapy*, Vol. 6. <https://doi.org/10.1186/s13195-014-0073-2>
- Le, W., Dong, J., Li, S., & Korczyn, A. D. (2017). Can Biomarkers Help the Early Diagnosis of Parkinson's Disease? *Neuroscience Bulletin*, Vol. 33, pp. 535–542. <https://doi.org/10.1007/s12264-017-0174-6>
- Lotankar, S., Prabhavalkar, K. S., & Bhatt, L. K. (2017). Biomarkers for Parkinson's Disease: Recent Advancement. *Neuroscience Bulletin*, Vol. 33, pp. 585–597. <https://doi.org/10.1007/s12264-017-0183-5>
- National Institute of Aging. (2018). What Is Lewy Body Dementia? Retrieved April 12, 2019, from <https://www.nia.nih.gov/health/what-lewy-body-dementia>
- Navlani, A. (2018). Decision Tree Classification in Python (article) - DataCamp. Retrieved April 15, 2019, from <https://www.datacamp.com/community/tutorials/decision-tree-classification-python>
- Orozco-Aroyave, J. R., Hönig, F., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Daqrouq, K., Skodda, S., ... Nöth, E. (2016). Automatic detection of Parkinson's disease in running speech spoken in three different languages. *The Journal of the Acoustical Society of America*, 139(1), 481–500. <https://doi.org/10.1121/1.4939739>
- Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), 1226–1238. <https://doi.org/10.1109/TPAMI.2005.159>
- Postuma, R. B., Iranzo, A., Hu, M., Högl, B., Boeve, B. F., Manni, R., ... Pelletier, A. (2019). Risk and predictors of dementia and parkinsonism in idiopathic REM sleep behaviour disorder: A multicentre study. *Brain*, 142(3), 744–759. <https://doi.org/10.1093/brain/awz030>
- Rektorova, I., Mekyska, J., Janousova, E., Kostalova, M., Eliasova, I., Mrackova, M., ... Marecek, R. (2016). Speech prosody impairment predicts cognitive decline in Parkinson's disease. *Parkinsonism and Related Disorders*, 29, 90–95. <https://doi.org/10.1016/j.parkreldis.2016.05.018>

- Rizek, P., Kumar, N., & Jog, M. S. (2016). An update on the diagnosis and treatment of Parkinson disease. *CMAJ*, Vol. 188, pp. 1157–1165. <https://doi.org/10.1503/cmaj.151179>
- Rocca, J. (2019). Ensemble methods: bagging, boosting and stacking – Towards Data Science. Retrieved April 25, 2019, from <https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205>
- Rusz, J., Cmejla, R., Ruzickova, H., & Ruzicka, E. (2011). Quantitative acoustic measurements for characterization of speech and voice disorders. *Journal of Acoustical Society of America*, 129(1), 350–367. <https://doi.org/10.1121/1.3514381>
- Rusz, J., Hlavnička, J., Tykalová, T., Bušková, J., Ulmanová, O., Růžička, E., & Šonka, K. (2016). Quantitative assessment of motor speech abnormalities in idiopathic rapid eye movement sleep behaviour disorder. *Sleep Medicine*, 19, 141–147. <https://doi.org/10.1016/j.sleep.2015.07.030>
- Sakar, B. E., Isenkul, M. E., Sakar, C. O., Sertbas, A., Gurgen, F., Delil, S., ... Kursun, O. (2013). Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics*, 17(4), 828–834. <https://doi.org/10.1109/JBHI.2013.2245674>
- Sanjeevi, M. (2017). Chapter 4: Decision Trees Algorithms – Deep Math Machine learning.ai – Medium. Retrieved April 15, 2019, from <https://medium.com/deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-b93975f7a1f1>
- Singh, G. (2006). Determination of Cutoff Score for a Diagnostic Test. In *The Internet Journal of Laboratory Medicine* (Vol. 2). Retrieved from <https://print.ispub.com/api/0/ispub-article/9884>
- Walker, Z., Possin, K. L., Boeve, B. F., & Aarsland, D. (2015). Non-Alzheimer's dementia 2: Lewy body dementias. *Lancet*, 386(10004), 1683–1697. [https://doi.org/10.1016/S0140-6736\(15\)00462-6](https://doi.org/10.1016/S0140-6736(15)00462-6).Non-Alzheimer
- Zhang, Y., Jiang, J., & Rahn, D. A. (2005). Studying vocal fold vibrations in Parkinson's disease with a nonlinear model. *Chaos*, 15(3). <https://doi.org/10.1063/1.1916186>