Explaining Parkinson’s Disease Computational Diagnostic based on Speech Analysis

Artur Oliveira Fortunato

Thesis to obtain the Master of Science Degree in

Computer Science and Engineering

Supervisors: Doctor David Manuel Martins de Matos
Doctor Alberto Abad Gareta

Examination Committee

Chairperson: Doctor name-of-president
Supervisor: Doctor David Manuel Martins de Matos
Member of the Committee: Doctor name-of-member-of-committee

September 2021

Agradecimentos

Magníficos Agradecimentos

Lisboa, September 22, 2021
Artur Oliveira Fortunato

Dedicatória interessante

Resumo

[TRADUZIR PARA PORTUGUÊS] Parkinson's Disease (PD) is a neurodegenerative disorder that affects the central nervous system. The disease manifests itself in the patient's speech, which usually becomes slurred, monotonic, and breathy. This symptoms provide a powerful biomarker for the detection of PD. The present work will analyse the subject's speech, representing it with a set of acoustic features. With this repre- sentation, a Machine Learning (ML) model will be trained to diagnose PD. By performing cross-language tests on this classification model, we will evaluate the hypothesis that a ML classifier can correctly diagnose PD. In addition, this diagnosis will be independent of the language spoken by the subject. Furthermore, we will explore the use of an explain- ability model, which will "translate" the classification model's diagnosis to a medical professional. The model will provide essential information for the clinician to trust and use this tool. Despite the good results achieved by many ML classification models, their acceptance for the diagnosis of this disease has not yet been achieved. Clinicians are unable to use the model's diagnosis, as it lacks a medical-oriented interpretation. Therefore, this work is motivated by the practical need of a universal, language-independent model that can provide enough human-understandable information to support clinical usage of ML models on PD diagnosis. This project will provide a tool that can boost the transfer of such models from test environments to real-life usage.

Abstract

Parkinson's Disease (PD) is a neurodegenerative disorder that affects the central nervous system. One of the disease's manifestation is in the patient's speech, which usually becomes slurred, monotonic, and breathy. These symptoms provide a powerful bio-marker for the detection of PD. The present work analyzes the subject's speech, representing it through a set of acoustic features. With this representation, a Machine Learning (ML) model will be trained to diagnose PD. By performing cross-language tests on this classification model, we will evaluate the hypothesis that a ML classifier can correctly diagnose PD. In addition, this diagnosis will be independent of the language spoken by the subject. Furthermore, we will explore the use of an explainability model, which will "translate" the classification model's diagnosis to a medical professional. The model will provide essential information for the clinician to trust and use this tool. Despite the good results achieved by many ML classification models, their acceptance for the diagnosis of this disease has not yet been achieved. Clinicians are unable to use the model's diagnosis, as it lacks a medical-oriented interpretation. Therefore, this work is motivated by the practical need of a universal, language-independent model that can provide enough human-understandable information to support clinical usage of ML models on PD diagnosis. This project will provide a tool that can boost the transfer of such models from test environments to real-life usage.

Palavras Chave
Keywords

Palavras Chave

Aprendizagem de máquina

Discurso

Explicabilidade

Interpretabilidade

Keywords

Machine Learning

Speech

Explainability

Interpretability

Table of Contents

List of Figures

List of Tables

1 Introduction

Neurodegenerative diseases are the most debilitating disorders that ail human kind, and the fourth leading cause of death. Neurodegenerative diseases affect the patient's thinking, movement, cognitive behavior, and memory, causing impairments and disabilities. These diseases include serious disorders like Alzheimer's Disease (AD) and Parkinson's Disease (PD) (Rai et al., 2019).

PD is the second most common neurodegenerative disease. It was estimated that 1% of people over 60 years old are affected with PD (Tysnes & Storstein, 2017). In 2015, more than 6 million people suffered from this disease worldwide. This value is projected to double by 2040, mainly driven by the increase of life expectancy (Dorsey, Sherer, Okun, & Bloem, 2018). One of the consequences of PD is the development of dementia. Almost half the PD patients develop dementia in the first 10 years after diagnosis (Williams-Gray et al., 2013), reaching over 80% after 20 years (Hely, Reid, Adena, Halliday, & Morris, 2008).

Early detection of PD can be critical for the life quality of the patient. Hence, the earlier the diagnostic is made, the earlier the treatment can begin, thus starting to control the evolution of the disease and improving the comfort of the patient. Furthermore, the majority of the treatment costs occur during the later stages of the disease, reinforcing the importance of early diagnosis (Pagan, 2012).

Over the last years, medicine and health care have been a prime focus for Artificial Intelligence (AI) and Machine Learning (ML). Numerous models have been tested to these areas, demonstrating impressive results in early detection of many diseases, among other tasks. However, the majority of these experiments focuses only on maximizing accuracy performance. Hence, a large problem remains unsolved on the real application of the previously referred models, as explainability has yet to become a focus for any of these works (Magesh, Myloth, & Rijo, 2020). Replacing medical decision-making with non-explainable, black-box ML models, can be contravening with the profound ethical responsibilities of clinicians (London, 2019). Consequently, the lack of explainability and interpretability of ML models used in these areas can seriously limit their chances of adoption in real practice (Vellido, 2019). Therefore, the application of explainable models will increase the possibility for medical professionals to understand a model's output, thus increasing the acceptance of AI systems in such tasks (Holzinger, Biemann, Pattichis, & Kell, 2017).

This work aims to improve our ability to automatically and more accurately diagnose PD, by building a classifier that differentiates PD patients from healthy people. This work aims, not only to allow PD to be detected earlier, but also to increase the degree of confidence of the diagnostic, leveraging the accuracy of the classifier close to 100%. Additionally, by using an explainability model, human-understandable explanations for the given classification (Parkinson or Healthy) of each patient will be generated to foster the use of ML models to support PD's diagnosis.

The document is structured as follows. Section 2 describes PD and state-of-the-art methodologies for PD computational diagnosis. Next, section 3 dives into the concept of Explainable Artificial Intelligence (XAI), and reviews multiple approaches developed in this area. Section 4 describes the experimental setup of this work, followed by section 5, which reports the results and their discussion. Finally, Section 7 presents the conclusions and future work.

2 Parkinson's Disease

PD is a common cause of dementia. It consists of a neurodegenerative disorder that affects the central nervous system. Symptoms begin gradually and worsen over time (National Institute of Aging, 2017).

Dementia is a syndrome that reduces the cognitive function. It affects a wide range of mental capacities, such as memory, comprehension, calculation, and language. Several factors can contribute to the development of dementia: factors that cannot be changed, such as age (the percentage of people over 60 years old suffering from dementia is estimated to be 5-8% (World Health Organization, 2020)) and genetics (having a family history of dementia increases one's probability of developing some form of dementia), and factors that can be changed, such as a poor diet and lack of exercise, smoking, diabetes, vitamin and nutritional deficiencies, excessive alcohol consumption and cardiovascular risk factors (Mayo Clinic, 2019).

Dementia impacts the patient physically and psychologically, but also its carers and family (as it creates a high level of dependency for the patient), and society at economic and social levels (World Health Organization, 2020).

2.1 Symptoms

As previously stated, symptoms begin gradually and worsen over time (National Institute of Aging, 2017). The most common symptoms include resting tremors (where hands or arms start shaking when resting), bradykinesia (or slowness of movement), muscle stiffness, which results in difficulty in moving and producing facial expressions, postural instability, which reduces the ability to maintain a steady posture, and dystonia, a condition in which patients have involuntary and repetitive muscle movements. In particular, PD also affects speech ability. Slurring and mumbling are observed in PD patients' speech, which is often observed to be monotone and breathy. The speech rate is also affected, as most patients speak slowly, although others speak too fast. Finally, cognitive problems have been associated with the disease, manifested as a difficulty in finding the correct words (which also contributes to slowing the speech) (Clinic, 2020).

2.2 Speech Impairments

PD patients exhibits multiple speech impairments, both at acoustic and at language levels. Acoustic parameters of speech, such as the Fundamental Frequency (F0) (Harel, Cannizzaro, Cohen, Reilly, & Snyder, 2004), pause duration (Harel et al., 2004) or vowel space time (Goberman & Elmer, 2005) have been shown to distinguish PD from Healthy Controls (HC).

90% of PD patients are reported to have speech and voice disorders (Froelich, Wróbel, & Porwik, 2015), which show that this biomarker can be an important source of information to detect PD. Instances of incomplete closure of vocal folds along with bowing folds during phonation have been reported (Perez, Ramig, Smith, & Dromey, 1996), leading to noise presence, typically characterized by measures such as Glottal-to-Noise Ratio (GNR), Noise-to-Harmonics Ratio (NHR), Harmonics-to-Noise Ratio (HNR) and Voice Turbulence Index (VTI). An increase in the average values of F0, jitter, and shimmer have also been measured in PD patients.

2.3 Diagnosis

To measure the progression of PD, Hoehn and Yahr (H&Y) was proposed (Hoehn & Yahr, 1967). This scale describes the disease through five stages, ranging from stage one, where the patient's symptoms are mild and only manifest on one side of the body, to stage five, where patients are confined to a wheelchair or a bed.

Two decades later, Unified Parkinson's Disease Rating Scale (UPDRS) was created (Fahn & Elton, 1987). It consists on a 50 question test, separated into four sections: (1) Mentation, behavior, and mood, (2) Activities of daily living (ADL), (3) Motor, and (4) Complications (Fish, 2011). Sections 2 and 3 include speech evaluation (section 2 contains questions for self-assessment of speech impairments, while section 3 evaluates free speech from the patient, for an evaluation made by the clinician). The UPDRS test has proven to be very effective in the diagnosis of PD, although favoring moderate and severe impairments. Hence, it may not be ideal to detect mild disease-related signs and symptoms (Goetz et al., 2003). Furthermore, the UPDRS presents some ambiguities (Goetz et al., 2003). In 2003, the Movement Disorder Society (MDS) published a state-of-the-art review on the UPDRS stating the test's inability to measure the severity of PD (Goetz et al., 2008). Furthermore, the heterogeneity of the test (evident on section 4 that uses a mixture of 5-point options with "yes" or "no" questions) hampers a global analysis. Additionally, some redundancy has been found in ADL and motor sections, which increases the time required to perform the test. Lastly, cultural biases have also been identified, in items such as Dressing and Cutting Food/Handling Utensils.

The MDS created the MDS-UPDRS (Goetz et al., 2008), an update on the UPDRS correcting several problems recognized in the original version. Specifically, the response type was homogenized (all questions are answered in a 4-point scale). Some of the redundant items were removed and unrepresented areas in the UPDRS were added. Finally, cultural bias was eliminated. An appendix regarding non-motor problems was also inserted in the MDS-UPDRS.

2.4 Computational diagnosis

Over the last years, many experiments have been conducted to diagnose PD using ML models. Such projects have achieved positive results, which are reviewed in this section.

2.4.1 Speech production tasks

The most common speech production tasks used for PD classification are:

- productions of a sustained vowel, as there are major variations in glottal noise and tremors in patients with PD (Godino-Llorente, Shattuck-Hufnage, Choi, Moro-Velázquez, & Gómez-García, 2017)
- Diadochokinesia (DDK), which consists of a fast repetition of sounds that imply quick succession of movements with the mouth and tongue (for this task, it is normal to use the pseudo-word */pa-ta-ka/*)
- Text-dependent Utterances (TDU)
- Text reading

Several speech production tasks to detect PD were tested (Pompilli et al., 2017) – Sustained vowel phonation (*/a/*), maximum phonation time (*/a/*), rapid repetitions of the pseudo-word */pa-ta-ka/*, reading of words, sentences and texts, and storytelling guided by visual stimuli. Two approaches were carried out. First, a sentence-level vector was created, with which the classifier achieved accuracies between 55% (with a sustained vowel phonation */a/* production task) and almost 71% (where the speech production task was reading out loud prosodic sentences). Secondly, all sentences were segmented into 4-second segments, with a time shift of 2 seconds. Using the features extracted at a segment level, the classifier achieved accuracies between 58% (with a sustained vowel phonation /a/ production task) and 85% (where the speech production task was reading of prosodic sentences). For this work, the authors used the FraLusoPark dataset (Pinto et al., 2016), which contains audio from 60 PD and 60 HC. The participants were European Portuguese speakers.

A set of 22 acoustic features was extracted from the Parkinson's Disease Detection Dataset (Little, McSharry, Costello, & Moroz, 2007) and the Parkinson's Telemonitoring Dataset (Tsanas, Little, McSharry, & Ramig, 2009). The Parkinson's Disease Detection Dataset includes speech by 23 patients with PD and 8 HC producing sustained vowels. The Parkinson's Telemonitoring Dataset contains speech from 42 PD patients producing sustained vowels. Using multiple ML classifiers, the system achieved an accuracy of almost 97% using a Gaussian Process Classification (GPC). With this model, the sensitivity reached 88% and the specificity went slightly above 97% (Despotovic, Skovranek, & Schommer, 2020).

To study the relevance of each phonemic group in detecting PD, three datasets were used – GITA (Orozco-Arroyave, Arias-Londoño, Vargas-Bonilla, Gonzalez-Rátiva, & Nöth, 2014), Neurovoz (Moro-Velazquez, Gomez-Garcia, Godino-Llorente, & Dehak, 2019), and CzechPD (Rusz et al., 2013). Neurovoz contains the results for multiple tasks – DDK, TDU and a monologue, based on a picture description – from 47 PD patients and 32 control Spanish Castilian speakers. GITA contains multiple speech production tasks from 50 PD patients and 50 HC Spanish Colombian speakers – DDK, TDU and a monologue. The CzechPD subset considered for this study contains only the DDK task, produced by 20 newly diagnosed and untreated speakers with PD and 14 HC, all Czech speakers. Using a Gaussian Mixture Model - Universal Background Model (GMM-UBM) classifier pre-trained with an auxiliary Spanish Castilian dataset, Albayzin (Moreno et al., 1993), the model yielded an classification accuracy of 94% for the CzechPD dataset, 89% for Neurovoz, and 84% for GITA (Moro-Velázquez et al., 2019).

Sustained vowels and text reading tasks were tested to differentiate PD from HC (Braga, Madureira, Coelho, & Ajith, 2019). The authors use three datasets – Proença (Proença et al., 2014) (containing audio from 22 PD patients in European Portuguese), UCI (Erdogdu Sakar et al., 2013) (with audio from 20 PD and 20 HC) and a dataset created for the purpose of this study by the authors. The Proença dataset contains word and text reading tasks and the UCI contains results from the sustained vowel task from the patients and healthy controls. The authors tested multiple ML classifiers, such as Neural Networks (NN), Support Vector Machines (SVM) and Random Forests (RF). This work yielded an accuracy of almost 95% with the RF classifier and slightly above 90% with NN (with 4 layers, comprising 7, 7, 6 and 7 neurons, respectively) and SVM.

2.4.2 Feature selection

Multiple acoustic features have been used to attempt to distinguish between PD and HC.

Cases of incomplete vocal folds closure along with folds bowing during phonation were reported (Perez et al., 1996), leading to the presence of noise, that is typically characterized using measures such as NHR, GNR, HNR, and VTI. Some feature values have also been found to increase in PD patients, such as average F0 and jitter (Bang, Min, Sohn, & Cho, 2013) and shimmer (Kent, Vorperian, Kent, & Duffy, 2003).

A set of 5 acoustic features – F0, correlation dimension, HNR, detrended fluctuation analysis and recurrence period density entropy – were selected from a set of 22 acoustic features by using Gaussian processes for regression and classification combined with Automatic Relevance Determination (ARD) (Despotovic et al., 2020). The authors tested multiple ML classifiers (SVM, RF, GPC, among others). The GPC achieved an accuracy of almost 97%, although the model's sensitivity was left on 88% (wrongly classifying 12% of the patients). The specificity reached 97%.

The adequacy of different phonemic groups in identifying PD patients was analyzed (Moro-Velázquez et al., 2019). The work describes the concept of phonemic grouping, which consists of grouping phonemes by their type (such as nasal, fricatives, plosives). Using a GMM-UBM classifier, this work yielded results with accuracies between 77% (using the plosive-nasal-vowel phonemic group) and 94% (with the fricative-nasal phonemic group). The authors extracted Rasta-Perceptual Linear Predictive (Rasta-PLP) (Hermansky, Morgan, Bayya, & Kohn, 1992) and its derivatives, $\Delta + \Delta\Delta$, and labeled them by phonemic group. The focus on the most important sounds has proved that plosive, vowel and fricative segments are the most important for PD detection.

A NN was trained with the VoxCeleb 1 (Chung, Nagrani, & Zisserman, 2017) and 2 (Chung, Nagrani, & Zisserman, 2018) datasets. An affine transformation was applied to the last pooling layer, to retrieve the *x-vectors*, an abstract representation of the input features, which were Mel-frequency cepstral coefficients (MFCC) and its derivatives, $\Delta + \Delta\Delta$. The *x-vectors* are then used as an input to a Probabilistic Linear Discriminant Analysis (PLDA) classifier. The model achieved an accuracy of 90% on TDU production tasks and 79% on DDK production task (repetition of the pseudo-word */pa-ta-ka/*) (Moro-Velázquez, Villalba, & Dehak, 2020).

2.4.3 Classification models

Most of the available datasets for this task are very small, considering the usual size for a classification problem. This property made the PD detection difficult. Indeed, complex models are unable to capture the variability of the data from a small dataset, and are therefore unable to correctly simulate and generalize the training set (Andonie, 2010). Therefore, the majority of the approaches to this problem use traditional machine learning models, such as SVM, RF and K-Nearest Neighbours (KNN), which are able to make accurate predictions training with small datasets. Nevertheless, some experiments have used Multi-Layer Perceptrons (MLP) and other NN architectures, achieving accurate results, in some cases yielding superior performances when compared to other models, such as SVM, and RF (Wan, Liang, Zhang, & Guizani, 2018).

A 114-dimensional feature vector was used as input to a RF. Using acoustic features such as F0, loudness, shimmer, jitter and MFCC, and using 5-fold cross-validation, the classifier achieved an accuracy of 85.1% (Pompilli et al., 2017).

A set of classifiers was used on two PD datasets (Despotovic et al., 2020). The authors extracted the top 5 acoustic features (using ARD) from a set of 22 features. After feature selection, the model achieved an accuracy of almost 97%, using a GPC with Matérns 3/2 and 5/2 as covariance functions. The SVM classifier yielded an accuracy close to 97% as well, whereas the Boosting Classifier (BC) obtained an accuracy around 1% lower, completing the task with close to 96 % accuracy. The RF achieved 96.62% specificity, whereas the model's accuracy almost reached 93%.

From the Naranjo dataset (Naranjo, Pérez, Campos-Roca, & Martín, 2016) 240 recordings were retrieved (Yaman, Ertam, & Tuncer, 2020). From these recordings, 44 acoustic features were extracted. The authors used KNN and SVM classifiers, achieving similar results, yielding accuracies slightly above 91%.

From the Naranjo dataset, a total of 177 acoustic features were retrieved (Yaman et al., 2020). Using the Relief algorithm, the authors selected the 66 more relevant features. Ensemble KNN was compared against Cosine KNN and Gaussian SVM was compared to Quadratic SVM. The Cosine KNN yielded an accuracy slightly above 91%, whereas the Gaussian SVM outperformed the Quadratic SVM, with an accuracy similar to the Cosine KNN (also above 91%).

A total of 2330 acoustic features were extracted from the mPower dataset (Bot et al., 2016) (2268 corresponding to Audio/Visual Emotion and Depression Recognition Challenge (AVEC) 2013 and 62 corresponding to GeMAPS) (Tracy, Özkancab, Atkins, & Ghomi, 2020). With 2023 HC and 246 PD, the authors tested three ML methods to distinguish between PD and HC: L2-regularized Logistic Regression (LR), RF, and gradient-boosted Decision Trees (DT). Because the dataset is heavily biased towards HC (n = 2023) compared to PD (n = 246), the authors added precision, recall and F1-score to the accuracy as evaluation metrics to compare the performance of each model. The gradient boosted DT achieved the best results, yielding 0.797 for recall, 0.901 precision and an F1-score of 0.836. Similar results were reached with the RF classifier, but with an inferior value for recall (0.693 recall, 0.902 precision and 0.783 for F1-score). The LR achieved the worst results, reaching 0.759 recall, 0.811 precision and 0.784 of F1-score.

A GMM-UBM classifier was trained using one dataset and tested it with three others. The model yielded accuracies between 84% and 94% (Moro-Velázquez et al., 2019).

MLP have also been extensively used for PD classification, having proven their efficacy in performing this task. A 1 hidden layer MLP, used on various sets of acoustic features, was able to classify AD patients with an accuracy of over 92% and HC with an accuracy of almost 91%, surpassing the performance of a KNN model, which yielded accuracies of 90.9% for AD and 87.3% for HC (Lopez-de Ipiña, Solé-Casals, et al., 2014). The Levenberg-Marquardt and Scaled Conjugate Gradient methods were tested as training algorithms for an MLP (Bakar, Tahir, & Yassin, 2010). Using 16 classical acoustic features (such as F0, jitter, shimmer) extracted from 195 speakers, the authors tested multiple values for the number of hidden units (5, 10, 15, 20, 25) and concluded that the Levenberg-Marquardt outperformed the Scaled Conjugate Gradient, reaching accuracies of over 97% with 25 hidden units, whereas Scaled Conjugate Gradient achieved 79% on 10 hidden units. Using the UCI dataset (Erdogdu Sakar et al., 2013), a set of 23 features was extracted for PD classification (Wan et al., 2018). The authors compared the performance of a Deep Multi-Layer Perceptrons (DMLP), with 5 or 10 hidden layers, with other ML classifiers. The authors reduced the size of the DMLP to 5 hidden layers, using *ReLU* or *softplus* as non-linear activation functions instead of the latter activation function, as these are continuous and can therefore address the vanishing gradient problem that affects Deep Neural Networks (DNN). Results on this experiment concluded that the best performance came from the DMLP using 10 hidden layers, which yielded 80% accuracy, whereas the LR model only reached 77.5% and the KNN could only get to 72.5%. Dropping the size of the DMLP to 5 hidden layers reduced the model's accuracy to 76%, which was still higher than some of the tested models, such as the KNN and RF models.

2.4.4 Universality

As the goal for this work is to develop a model capable of detecting PD for any patient, universality is an important property for the desired model, which can be achieved with language-independency.

Three distinct datasets, one in Spanish, one in German and one in Czech, were used with a GMM-UBM model to train a semi language-independent model (Orozco-Arroyave et al., 2016). For each experiment, the model was trained with one dataset and tested with another (adding to the training set subsets of the test set with percentages varying from 10% and 80%). Despite reaching accuracies of 96%, high accuracies are only achieved when large portions of the test language are used to train the model. In a fully language-independent model (where the model is trained using one language and tested with another), the model accuracy only reaches 77% (trained with the German dataset and tested with the Czech dataset).

A GMM-UBM was trained using *corpora* in Spanish Castilian, Spanish Colombian and Czech. Cross-language testing resulted in accuracies of 82% (Moro-Velázquez et al., 2019).

3 Explainability Models

XAI is a field of AI that provides techniques and algorithms able to generate interpretable, intuitive, human-understandable explanations of AI decisions (Das & Rad, 2020).

Explaining the decisions made by a black-box model requires knowledge of its internal operations (Das & Rad, 2020), which makes it impossible to use by end-users who are only focused and interested on getting an accurate result. The very nature of a *black-box* ML/Deep Learning (DL) model is a barrier for their real-life usage (Shrikumar, Greenside, & Kundaje, 2017). For a ML model be used in real life situations, the users must have confidence in it. Two definitions of *trust* must be considered: *trust in the prediction*, where the user trusts a prediction sufficiently such that he is comfortable with performing an action based on it, and *trust in the model*, which gives enough confidence to deploy the model. Thus, in order for such model to be deployed, both definitions must be fulfilled (Ribeiro, Singh, & Guestrin, 2016). This is even more important in critical situations, such as medical diagnosis. To address this limitation in ML and DL, many models have been created to generate explanations for a model's predictions.

Creating human-understandable explanations can also aid in finding erroneous behavior in a model. A peculiar discovery was made in an experiment where Fisher Vector classifiers were used for the image recognition task (Bach, Binder, Montavon, Müller, & Samek, 2016). An interpretability technique called Layer-wise Relevance Propagation (LRP) was applied to explain the predictions of the model. In particular cases, where the input image consisted of a horse, it was found that the model primarily based its decision not on any of the physical traits of the horse, but on a copyright tag present on the bottom left of the image that turned out to be a characteristic of all the horse images used in training. This error certainly highlights the need for interpretability of ML/DL models, especially in the medical field, where such errors can severely impact human lives.

3.1 Explanation

An explanation is a verifiable justification for a model's output or decision (Das & Rad, 2020). There are many kinds of explanations, such as a heat map stressing relevant parts of an image (for example, a DaTscan image in PD detection (Magesh et al., 2020)). Some models, such as Local Interpretable Model-agnostic Explanation (LIME) (Ribeiro et al., 2016), base their explanations on activations or parameters of the black-box models, using simpler surrogate models (Das & Rad, 2020).

3.2 Scope

Explainability models can be subdivided in three large groups, based on the scope of their explanations: local, global or mixed.

3.2.1 Local explanations' models

Locally explainable methods are designed to generate an explanation for the model's decision on a single instance of input data (Das & Rad, 2020). Models that provide local explanations fail to provide a global observation of the model. Their explanations do not provide enough information on the original model computations and do not provide enough detail to understand the model's behavior as a whole (Agarwal, Frosst, Zhang, Caruana, & Hinton, 2020).

The concept of Axiomatic Attributions was proposed (Sundararajan, Taly, & Yan, 2017). Consider a function $F : \mathbb{R}^n \rightarrow [0, 1]$ representing a DNN. Let $x \in \mathbb{R}^n$ be the input, and $x' \in \mathbb{R}^n$ be the baseline input (the black image for image networks, for example). Using a straight line path in $\mathbb{R}^n$ from x' to x, the model computes the gradients along the path, in every point. Integrated gradients are obtained by cumulating these gradients. Specifically, integrated gradients are defined from baseline x' to input x as the path integral of the gradients along a straight line path. For each dimension i, $\frac{\partial F(x)}{\partial x_i}$ defines the gradient along dimension i. integrated Gradients (IG) are then calculated as

$$IG_i(x) = (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \tag{3.1}$$

The IG provide a measure of the relative importance of each feature on the model's classification of instance x - the higher the IG of feature i, the higher its importance for the classification.

Randomized Input Sampling for Explanation (RISE) was proposed in 2018. This model is based on random masking to locally understand the most important features (for example, in the case of the image classification problem, RISE will determine the most important pixels for the black-box model's classification) (Petsiuk, Das, & Saenko, 2018).

Consider $f : I \rightarrow \mathbb{R}$ to be the model. For the image classification problem, we consider $\Lambda : \{1, .., J\}$ x $\{1, .., W\}$ as the image coordinates and I would map every pixel to its RGB representation ($I = \{I | I : \Lambda \rightarrow \mathbb{R}\}$). f is a classifier that returns the probability of an instance of a certain class be present in the image. Considering a random binary mask $M : \Lambda \rightarrow \{0, 1\}$ following a distribution $\mathcal{D}$. By masking the image with $I \odot M$ (where $\odot$ represents the element-wise multiplication), we preserve only a subset of the pixels of I. By calculating the confidence score $f(I \odot M)$, we can define the importance of every pixel λ, $S_{I,f}(\lambda), \lambda \in \Lambda$, as the average value of the confidence scores of all masked images where $M(\lambda) = 1$. Mathematically,

$$S_{I,f}(\lambda) = \frac{1}{E[M]} \sum_{m \in M} f(I \odot m) \cdot m(\lambda) \cdot P[M = m] \tag{3.2}$$

3.2.2 Global explanations' models

Understanding the model's behavior on a set of input data points could provide insights on the input features, patterns, and their output correlations, thereby providing transparency of model behavior globally. Various globally explainable methods break down complex deep models into linear counterparts, which are easier to interpret (Das & Rad, 2020).

To generate explanations at class , Concept Activation Vectors (CAV) were proposed, which provide interpretations for a NN's internal state in terms of human-friendly concepts (Kim et al., 2018). The model considers a NN with inputs $x \in \mathbb{R}^n$ and a feed-forward layer l with m neurons. Thus, layer l's activation can be seen as $f_l : \mathbb{R}^n \rightarrow \mathbb{R}^m$. The user chooses a concept of interest C and creates a series of inputs llevelabeled as *contains concept* C and a series of inputs labeled as *does not contains concept* C. The model then calculates the hyperplane separating the two groups of inputs. The CAV is then defined as the vector normal to this hyperplane.

Common interpretability methods (such as saliency maps), calculate the derivatives for the *logit* in terms of the input features. With this approach, these methods are able to measure sensitivity in the set of input features. When combining CAVs and directional derivatives, the model can gauge sensitivity of ML predictions in directional input changes of the concept C, at activation layer l.

In 2020, the concept of Neural Additive Models (NAM) was proposed (Agarwal et al., 2020). The explanations are created by shape functions, relative to each input feature. To parameterize the these functions, a NN is created for each function. With this architecture, the model is able to create an exact representation of how NAMs compute a prediction, thus creating an explanation of the model's global behavior.

Consider $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ as the training set, with N instances, where x is the input vector and y is the target vector. The proposed model was trained using the following loss function:

$$\mathcal{L}(\Theta) = E_{x,y\sim\mathcal{D}}[l(x,y;\Theta) + \lambda_1\eta(x;\theta)] + \lambda_2\gamma(\Theta) \tag{3.3}$$

where $\eta(x,\Theta) = \frac{1}{K}\sum_x\sum_k(f_k^\Theta(x_k))^2$ is the output penalty, $\gamma(\Theta)$ is the weighted decay and f_k^Θ represents the k^{th} feature network.

The authors use the cross-entropy loss for binary classification as the task-dependent loss function $l(x,y;\Theta)$, which, considering $p_\Theta(x) = \sigma(\beta^\Theta + \sum_{k=1}^{K} k_k^\beta(x_k))$, yields

$$l(x,y;\Theta) = (\beta^\Theta + \sum_{k=1}^{K} f_k^\Theta(x_k) - y)^2 \tag{3.4}$$

where β^Θ defines the parameters to be calculated.

3.2.3 Mixed models

To combine the advantages of the local and global explanations' models, mixed models provide explanations that are able to locally interpret decisions, while also allowing to understand the behavior of the model as a whole.

Similarly to RISE, LRP allow to understand which pixels of the image contribute the most to the model's decision (Lapuschkin et al., 2015). This model, created for DNN architectures, redistributes the relevance of each neuron at the last layer of the network to pixel-wise scores (R_i^l) using the rule

$$R_i^{(l)} = \sum_j \frac{z_{ij} \sum_{i'} z_{i'j}}{R}{}_j^{(l+1)}, z_{ij} = x_i^{(l)} w_{ij}^{(l,l+1)} \tag{3.5}$$

where i is the i-*th* neuron in layer l, $\sum_j$ iterates through all the upper-layer neurons to which neuron i contributes. This result can be assessed using a visualization tool, such as a heat map, to explain the model's classification.

LIME is an algorithm that uses *local interpretable representations* of the classification data to generate an output that can be interpreted by humans (Ribeiro et al., 2016). We define $x \in \mathbb{R}^d$ as the original representation of the instance to be explained and $x' \in \{0,1\}^{d'}$, a binary vector and its interpretable representation. Let $g \in G$, where G is the set of models that can present a interpretable output to the user. We also denote $\Omega(g)$ as a measure g's explanation complexity and $f : \mathbb{R}^d \to \mathbb{R}$ as the model to be interpreted. $f(x)$ will be the probability or binary indicator that x belongs to a particular class. Let $\pi_x(z)$ be a proximity measure of distance between x and an instance z to define around x. Lastly, we define $\mathcal{L}(f, g, \pi_x)$ as a measure of how unfaithful g is approximating f in the space defined by π_x. As we want to maximize interpretability while keeping local fidelity, the explanation can be defined as:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \, \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{3.6}$$

The algorithm creates samples around x', weighted by π_x. Considering a perturbed sample $z' \in \{0,1\}^{d'}$ containing a fraction of the non-zero elements of x', the original representation $z \in \mathbb{R}^d$ is obtained, so the value $f(z)$ can be calculated. Considering $\mathcal{Z}$ as the set of all perturbed z' with the label $f(z)$, equation 3.6 is used to calculate the explanation.

DeepLIFT (DeepLIFT) was presented as a method to understand the output of a NN by backpropagating the neurons' contributions to every feature of the input (Shrikumar et al., 2017). To assign contribution scores C, DeepLIFT compares the activation of each neuron t to its reference activation value t^0, using the summation-to-delta property:

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta t} = \Delta t \tag{3.7}$$

where $\Delta t = t - t_0$ and $C_{\Delta x_i \Delta t}$ is a measure of the difference from the reference value t^0 attributed to the neuron x_i. The reference values t^0 are calculated by defining a set of reference input values $x_1^0, x_2^0, ..., x_n^0$ for a given neuron, resulting in $t^0 = f(x_1^0, x_2^0, ..., x_n^0)$, where f is the activation function of the neuron. The choice of reference input values is highly context-dependent. For example, for the MNIST (LeCun & Cortes, 2010), the reference values were set to 0 (which represents the background color of the images, black). For DNA classification tasks, the references were defined based on the expected frequency of each of the elements on the DNA's alphabet (A, C, G, T). This creates a limitation on some applications, as defining reference values may be difficult.

3.3 Parkinson's Disease diagnosis

As stated in section 1, ML models used for sensitive tasks, such as detection of PD, lack the ability to generate an explanation to be interpreted by the medical professionals that need to establish a diagnosis. These models, called black-box models (Holzinger et al., 2017), take an input and return as an output a classification, which cannot be interpreted by a medical professional. This problem difficults the acceptance of these models for such tasks, as the risk of decision-making based on the results of a black-box system raises numerous ethical concerns (Chen et al., 2020).

Image-based explanations were generated for a black-box model (the VGG16 convolutional neural network) on a dataset of SPECT DaTSCAN images of the brain (Magesh et al., 2020). The authors retrieved a 2-dimensional section of the 3-dimensional image, trained, and tested the *black-box* model, which yielded an accuracy of 95.2%, a specificity of almost 91%, a sensitivity of 97.5% and a precision of 95.2%. After the classification, the authors generated a color map over the input images to highlight the regions of interest (the pixels with larger weights for the classification process). This showed that the most interesting regions of the brain for this task were the *putamen* and the *caudate*, confirming the medical background information described, providing trust in the model, as it could be easily interpreted by a medical professional.

Explainability models have been applied to many other medical tasks, such as breast cancer detection (Pfob et al., 2020), identification of individuals with high-risk of depressive disorder (Choi, Shim, Jeong, & Jo, 2020), and early detection of COVID-19 (Punn & Agarwal, 2020).

This area remains almost unexplored for the task of early detection of PD and, to the best of our knowledge, no work has combined explainability algorithms with acoustic-based models for this task.

4 Experimental Setup

This section describes the methodology. First, the *corpora* used in this work are described, followed by the approaches to be followed (feature selection, classification model, explanation generation model, and multi-language tests). Finally, the evaluation procedures are presented. Figure 4.1 shows the pipeline for the system's architecture.

4.1 Corpus Description

As described in section 3.3, most datasets available for this task are insufficient to train neural models (Andonie, 2010). Nevertheless, few common speech production tasks are available in the datasets. Because some of these datasets only differ in the model of the microphone used for recording and the language of the test subjects, several datasets may be combined to produce sufficiently long collections of data that can be used for neural models (Braga et al., 2019), (Despotovic et al., 2020), (Moro-Velázquez et al., 2019), (Moro-Velázquez et al., 2020). Different datasets were used for training and testing, or to combine instances from different datasets in the training and/or testing sets (Orozco-Arroyave et al., 2016), all proving to be accurate in the PD classification task.

This study used 3 datasets for training and testing the model – FraLusoPark (Pinto et al., 2016), GITA (Orozco-Arroyave et al., 2014), and Mobile Device Voice Recordings at King's College London (MDVR_KCL) (Jaeger, Trivedi, & Stadtschnitzer, 2019).

The FraLusoPark dataset is composed by speech from 120 patients, half of which are native French speakers and the other half are European Portuguese speakers. The dataset also contains 120 healthy participants as a control group (with the same distribution between French and European Portuguese speakers as the PD participants). Each group of PD patients is divided into three subgroups, based on the number of years since diagnostic: 20 early stage patients (who have been diagnosed less than 3 years before and present no motor fluctuations), 20 mid stage patients (with a diagnostic made 4 to 9 years before the data collection, or less than 3 years and experiencing motor fluctuations), and 20 advanced stage patients, diagnosed over 10 years ago. The patients' speech is recorded twice for every speech production task, *before* (at least 12 hours after medication) and *after* medication (at least 1 hour after medication). FraLusoPark participants were asked to perform a set of speech production tasks:

- sustain the vowel *a* at a steady pitch
- hold a vowel during their maximum phonation time (*a*) on a single breath
- DDK (repetition of the pseudo-word *pa-ta-ka* at a rapid pace during 30 seconds)

- reading aloud 10 words and 10 sentences, formed by adapting part of section V.2 of the Frenchay Dysarthria Assessment of Intelligibility (FDA-2)
- reading of a short text (adapted to French and European Portuguese)
- storytelling by guided visual stimuli
- reading a collection of sentences with specific language-dependent prosodic properties
- free conversation for 3 minutes

In the scope of the present study, we will only consider the Portuguese speakers of this dataset, as the audios from the french patients could not be accessed. These recordings total 62 minutes and 45 seconds for PD patients and 51 minutes and 2 seconds for HC participants.

The GITA dataset contains recordings of 50 PD patients and 50 HC, evenly distributed between genders. For the PD group, the average age is 62.2 with a standard deviation of 11.2 years and 60.1 with a standard deviation of 7.8 for male and female participants, respectively. Considering the HC group, the average age is 61.2 and 11.3 years and 60.7 with a standard deviation of 7.7 for male and female participants, respectively. Multiple stages of disease progression are considered in this study (time since diagnostic ranges between 0.4 - 20 years for male patients and 1 - 41 years for female patients). All the participants are Colombian Spanish native speakers. Recordings of the PD patients were made no up to 3 hours after the morning medication. Different speech production tasks were performed to examine phonation, articulation and prosody. To analyze phonation, participants were asked to sustain the five Spanish vowels and to repeat the same five vowels, but alternating the tone between low and high. Regarding articulation, a DDK evaluation was performed with the pseudo-words */pa-ta-ka/*, */pa-ka-ta/* and */pe-ta-ka/*. Finally, for the evaluation of prosody, both PD patients and HC were asked to repeat a series of sentences with different levels of complexity, to read a dialogue between a doctor and a patient, which contained the complete set of Spanish sounds, to read sentences with a strong emphasis on a set of words and freely speak about their daily routine. These recordings total 15 minutes and 31 seconds for PD patients and 14 minutes and 41 seconds for HC participants.

Lastly, the MDVR_KCL dataset was recorded in the context of phone calls, in an acoustically-controlled environment. The dataset contains speed from 16 participants with PD (11 male and 4 female) and 21 HC (3 male and 18 female), totaling 37 native English speakers. The PD group contains patients from all the stages of the disease (early, mid and late stages) according to the Hoehn and Yahr scale (Hoehn & Yahr, 1967). The participants were asked to read a text ("The north wind and the sun" or "Tech. Engin. Computer applications in geography snippet"). Additionally, the interviewer started a spontaneous conversation with each participant about various topics.

To homogenize the datasets, only the text-reading tasks were considered herein. This yields a total of 131 HC and 125 PD speakers of European Portuguese, Colombian Spanish, and European English.

4.2 Data Processing

The original audio files contained full interviews of each test subject, therefore requiring segmentation in order to remove useless audio fragments. Silences between speech segments were removed. Next, sounds produced by the subject that were not considered as speech were also removed. Finally, audio segments containing speech from interviewers were also eliminated.
After data processing, the data sets were reduced as follows for PD patients and HC participants:

- FraLusoPark recordings: 113 minutes and 43 seconds
- Gita recordings: 30 minutes and 8 seconds
- MDVR_KCL recordings: 65 minutes and 3 seconds

4.3 Feature Extraction

In order to extract the features, the openSMILE (Eyben, Wöllmer, & Schuller, 2010) tool was used. To extract the complete set of features, four configurations were used: *MFCC12_0_D_A.conf* for MFCC's, *PLP_0_D_A.conf* to extract Perceptual Linear Predictive (PLP)'s and *prosodyAcf2.conf*, for prosody features (F0 and HNR).

OpenSMILE was configured to use a sliding window of 25*ms* with a frame step of 10*ms*. After the extraction, each participant was represented by a list of frames, and each frame described by a list of features. To classify each patient, the resulting diagnostic is obtained by averaging the model's output for each of the patient's frames.

4.4 Classification Experiments

Three distinct experiments were conducted during the present work. First, a baseline was created by training and testing a classification model with sets from the same database. This procedure scored the classification models for single languages. Secondly, the same model was trained to evaluate its performance as a semi language-independent classifier. For this, the model was trained using one complete dataset and with a fraction of another dataset (this evaluation was achieved by training the model with a combination of one complete dataset with 90% of another dataset), thus combining two languages in the same training set. The model was then tested with the remaining 10% of the second dataset. All the combinations between the three datasets were tested, leading to 6 dataset combinations. By testing this semi language-independent version, it was possible to evaluate an intermediate step between a language-dependent and a language-independent classification model, shedding light into the model's sensitivity to the language. Lastly, a completely language-independent model was trained by combining two datasets. For this last experiment, each model was trained with two datasets and tested with the third, thus allowing to evaluate the model's ability to diagnose a patient who speaks in a language different from the ones used to train the model.

These experiments used the *scikit-learn* implementation (Pedregosa et al., 2011) of a MLP was used. Two different architectures were tested to evaluate their ability to learn from the training

data. The first architecture contains one entry layer with **N** neurons (where **N** is the number of input features), a fully-connected hidden layer with **N + 1** neurons and an output layer with 1 neuron, whose value represents the probability of the test subject to be classified as PD (Lopez-de Ipiña, Solé-Casals, et al., 2014). The second architecture also contains an input layer with **N** neurons, two fully-connected hidden layers, comprising 200 neurons each and, similarly to the first architecture, an output layer with 1 neuron, also representing the probability of the subject under evaluation to be diagnosed with PD. For these experiments, the threshold between HC and PD diagnostics (the output value of the neuron from the output layer) was set to 0.5.
In order to find the best model configuration, the experiments were repeated testing multiple values for the L2 regularization term parameter, or *alpha* (10^{-3}, 10^{-2}, 10^{-1}), maximum number of iterations (1000, 2000, 5000) and solver for weight optimization (lbfgs, sgd, adam).

4.5 Explanation Production

After the classification experiments, explanations were generated for each individual of the test set (with all models described in section 4.4). As the objective of this work is to generate an explanation for each diagnostic individually, a mixed model was used. By selecting a mixed model, this work can be extended to generate global explanations, thus shedding light on the role of each acoustic feature on the model's diagnostic.
The selected model was LIME (Ribeiro et al., 2016). This model yielded results on explaining PD diagnostics with SPECT DaTSCAN images of the brain that were confirmed by the bibliography. This work aimed to verify if a similar performance can be achieved using acoustic features.
To explain the diagnostic of each subject, the *explain_instance* method from the *LimeTabularExplainer* class was used with each feature list (each representing a time frame) to generate a report. Next, two operations were performed. First, the classification model's output was averaged between all time frames, creating a final classification probability for the subject. Secondly, each feature weight (which LIME calculated) was also averaged, thus creating a final weight for each feature (from which the top five features with the largest contribution to the classification were selected). To this report, a list of normal values for healthy patients for each of the features was added. Finally, to assist the interpretation of the report by the medical professional, a small description of each feature was also added. Tables 1 and 2 show the complete list of normal values and descriptions for each feature.

4.6 Model Evaluation

To evaluate the classification model's performance, multiple metrics have been selected:

- *accuracy* allows to evaluate the % of subjects correctly diagnosed
- *precision* yields the fraction of subjects diagnosed with PD that were correctly classified
- *recall* quantifies the percentage of PD subjects that were correctly diagnosed
- *F1-score* allows to evaluate *precision* and *recall* in the same metric
- *Specificity* is the fraction of subjects classified as HC that were correctly diagnosed

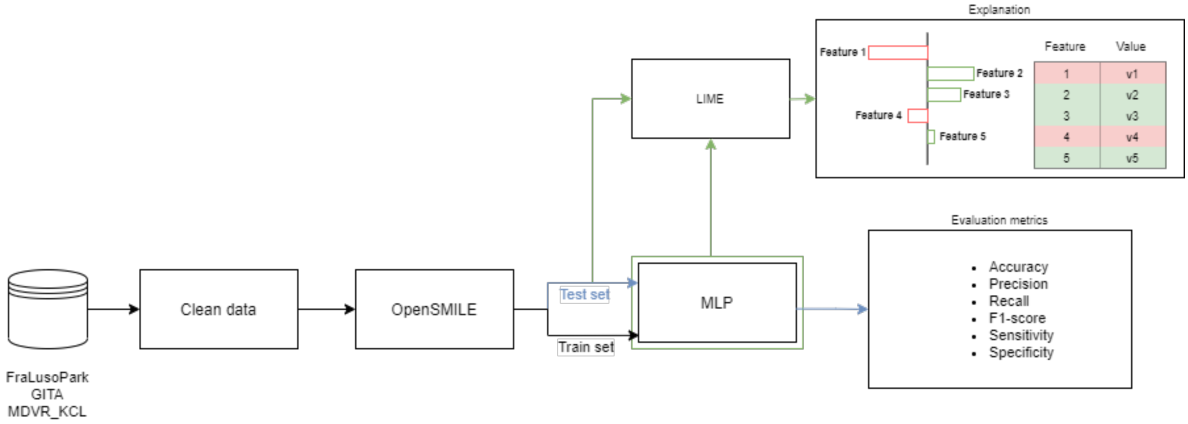


Figure 4.1: Pipeline of the proposed model.

These metrics quantify the performance of the models, which allows to determine the best parameters and architecture. Furthermore, *recall* allows to evaluate the percentage of subjects from the PD group that were correctly diagnosed, which, combined with specificity (that evaluates the number of subjects from the PD group incorrectly diagnosed), provides confidence in the model information to medical professionals.
To assess LIME's performance, average values of each feature were extracted from the bibliography (see **text**) and are shown along with the values in each explanation, in order to compare each subject's feature values with it's range for a healthy individual. This comparison will allow to evaluate the model's ability to detect abnormal values (or their absence) and select those features as justifications for a given classification.

5 Results and discussion

5.1 Classification Experiments

In this work, three types of experiments were conducted, each using two different architectures, as described in the previous chapter. Results are shown in tables **5.1** and **5.2** (for the baseline experiments), **5.3** and **5.4** (for the semi-independent experiments), and **5.5** and **5.6** (for the language-independent experiments). These tables show the five MLP parameter parameterizations with higher accuracy for each experiment. Tables **5.1**, **5.3** and **5.5** present the results for architecture 1, whereas tables **5.2**, **5.4** and **5.6** show the results for architecture 2.

5.1.1 Baseline experiments

Both architectures 1 and 2 of the MLP yielded an accuracy of 90% with the best parameterization (Tables **5.1** and **5.2**).

All the best models parameterizations (for both architectures 1 and 2) achieved higher scores using the GITA dataset. There are multiple reasons that can explain these results. In particular , the text read by subjects for the creation of the Gita dataset contains the complete set of Spanish sounds, which makes the data phonetically complete. Also, the audios from the MDVR_KCL dataset were recorded using phone calls, which uses audio compression with data loss, resulting in a dataset with inferior quality. In addition, MDVR_KCL has a significantly smaller recording time, which may limit the model learning.

The distribution between MLP solvers (adam and lbfgs) on the top 5 model parameterizations for architecture 1 is similar, whereas 4 out of the 5 best model parameterizations on architecture 2 use the adam solver. Both architectures yielded better results when using valuer smaller values (0.0001 and 0.001) for the alpha parameter, comparing to the results obtained using larger values (0.1). Finally, architecture 1 does not show significant differences between models using 2000 and 5000 for the maximum number of iterations. In addition, this difference is observable on architecture 2, where the four model configurations which yielded better results by using the value of 5000 for this parameter regardless of the solver. The difference between architectures can be explained by the higher complexity of architecture 2 which require the optimization of a large number of parameters (52400 weights and 401 biases), compared with architecture 1, which has only 3844 weights and 63 bias. A larger number of parameters requires more iterations for the model's convergence.

Architecture 1 yielded precision values between 0.75 and 1, meaning that 75% to 100% of the patients labeled as PD by the models were correctly classified. The precision of architecture 2 was slightly worse, between 67% and 100%. Recall values (which corresponds to the percentage of PD patients were correctly classified) were similar for the two architectures. Architecture 1 led to recall values in the ranges [71-100]% and [67-100]%, respectively. Using the specificity metric (which corresponds to the percentage of HC patients that were correctly classified) to

dataset	solver	alpha	max. iterations	accuracy	precision	recall	specificity	f1-score
gita	adam	0.0001	5000	0.9	0.75	1.0	0.857	0.857
gita	lbfgs	0.0001	2000	0.9	0.75	1.0	0.857	0.857
gita	adam	0.001	2000	0.8	1.0	0.75	1.0	0.857
gita	lbfgs	0.01	5000	0.8	0.833	0.833	0.75	0.833
gita	lbfgs	0.01	2000	0.8	1.0	0.714	1.0	0.833

Table 5.1: Baseline experiment results using architecture 1.

dataset	solver	alpha	max. iterations	accuracy	precision	recall	specificity	f1-score
gita	adam	0.001	5000	0.9	0.8	1.0	0.833	0.889
gita	lbfgs	0.001	5000	0.9	1.0	0.833	1.0	0.909
gita	adam	0.0001	5000	0.8	0.8	0.8	0.8	0.8
gita	adam	0.01	5000	0.8	0.667	0.667	0.857	0.667
gita	adam	0.001	2000	0.8	1.0	0.778	1.0	0.875

Table 5.2: Baseline experiment result using architecture 2.

compare the two architectures, architecture 2 outperformed architecture 1 by a small margin, producing a range of values between 80% and 100%, whereas architecture 1 produced a range of values between 75% and 100%. Finally, comparing both architectures using the F1-score metric, the performance of architecture 2 (up to 91%) is usually higher than the one of architecture 2 (up to almost 86%).
Overall, we can conclude that there are no significant differences between the two architectures.

5.1.2 Semi-independent experiments

When testing a semi-independent approach, architecture 1 yielded better results than architecture 2 (tables **table 5.3** and **table 5.4**). Although the two best model parameterization of both architectures produced an accuracy of 90%, the following three model parameterization resulted in an accuracy of almost 86%, whereas architecture 2 only reached an accuracy of 80%. The same trend applies to precision.
Architecture 1 outperformed architecture 2 on precision, producing results between 0.83 and 1, whereas architecture 2 yielded values between 0.6 and 1. While both architectures' highest value was the same, architecture 1 produced consistently better results, with a smaller range of values. Similar results were achieved when using recall. Architecture 1 produced values between 0.75 and 1, and 3 of the top 5 model parameterizations achieved 100% recall. Additionally, architecture 2 values for recall ranged from 0.66 and 1. As F1-score combines the values from precision and recall (and architecture 1 outperformed architecture 2 on both these metrics), the F1-score metric leads to the same conclusions. Values of this metric for architecture 1 varied between 0.85 and 0.92, whereas architecture 2 values ranged from 0.75 to 0.88. Finally, architecture 2 produced better results when using specificity. This architecture's values varied between 0.71 and 1, with a much smaller variation between extremes when compared to the results produced by architecture 1, which varied from 0.5 to 1. The results were similar to the ones achieved on the baseline experiences using architecture 2. Architecture 1 had a slightly better performance on the semi language-independent experiments, compared to the baselines. This experiment confirms the conclusions of similar work that tested semi language-independent models (Orozco-Arroyave et al., 2016), which suggests that these models can be

dataset	solver	alpha	max. iterations	accuracy	precision	recall	specificity	f1-scor
mdvr_kcl + gita	adam	0.001	2000	0.9	0.857	1.0	0.75	0.923
fralusopark + gita	lbfgs	0.0001	5000	0.9	0.875	1.0	0.667	0.933
gita + fralusopark	adam	0.0001	2000	0.857	0.833	1.0	0.5	0.909
gita + fralusopark	adam	0.01	5000	0.857	0.889	0.889	0.8	0.889
gita + fralusopark	lbfgs	0.001	2000	0.857	1.0	0.75	1.0	0.857

Table 5.3: Semi-independent experiment result using architecture 1.

dataset	solver	alpha	max. iterations	accuracy	precision	recall	specificity	f1-scor
mdvr_kcl + gita	lbfgs	0.001	5000	0.9	1.0	0.8	1.0	0.889
fralusopark + gita	lbfgs	0.0001	2000	0.8	0.75	0.75	0.833	0.75
mdvr_kcl + gita	adam	0.0001	5000	0.8	1.0	0.667	1.0	0.8
mdvr_kcl + gita	adam	0.001	5000	0.8	0.6	1.0	0.714	0.75
mdvr_kcl + gita	lbfgs	0.0001	5000	0.8	0.6	1.0	0.714	0.75

Table 5.4: Semi independent experiment result using architecture 2.

retrained using a small dataset of a new language. These retrained models can be used on patients that speak the different language, without loss of performance. This characteristic can be particularly useful, as lack of training data is usually a limitation to train such models.

5.1.3 Language-independent experiments

Language-independent models lead to substantially worse results compared to previous models (tables **table 5.5** and **table 5.6**).
When using a language-independent model, architecture 1 achieved a maximum accuracy of 67%. Architecture 2 yielded very similar results, scoring a maximum of 66% on this metric.
Combining the top five model parameterizations for both architectures, almost all (90%) obtained their best scores when trained with the FraLusoPark and MDVR_KCL, and tested with Gita. The same percentage of the combination of the top five models of each architecture used the *lbfgs* solver, whereas only 1 of these 10 model parameterizations used the *adam* solver. Similarly to the dependent and semi-independent experiments, the model's performance is consistently higher for smaller values of alpha. On both architectures, only 1 of the top five model parameterizations used $alpha = 1$. Finally, no significant differences were found when comparing model's performance based on the number of iterations.
Considering the precision metric, architecture 1 scored slightly higher values than architecture 2. It's values range between 0.59 and 0.64 whereas architecture 2 yielded values between 0.57 and 0.61, meaning that architecture 2 produced more false positives (patients from the HC group incorrectly classified as PD). Also, architecture 1 performed slightly worse when comparing the recall metric, only achieving values ranging from 0.76 to 0.84, whereas architecture 2 scored recall values between 0.77 and 0.88, thus correctly classifying a higher number of patients from the PD group. Architecture 1 outperformed architecture 2, when compared using the specificity metric. Architecture 2 only achieved a maximum of 0.46, compared to architecture 1, which scored a maximum of 0.58 on this metric. Lastly, as F1-score combines precision and recall in the same metric, the results of both architectures on this metric were equivalent.
We can conclude that the models have a similar performance on the PD detection task. Thus, Architecture 1 can be considered a better option for this task, as it is simpler, with only 3783

dataset	solver	alpha	max. iterations	accuracy	precision	recall	specificity	f1-score
gita	lbfgs	0.001	5000	0.67	0.644	0.76	0.58	0.697
gita	lbfgs	0.01	2000	0.65	0.612	0.82	0.48	0.701
gita	lbfgs	0.001	2000	0.65	0.615	0.8	0.5	0.696
gita	lbfgs	0.0001	2000	0.63	0.592	0.84	0.42	0.694
gita	adam	0.0001	5000	0.63	0.6	0.78	0.48	0.678

Table 5.5: Independent experiment result using architecture 1.

dataset	solver	alpha	max. iterations	accuracy	precision	recall	specificity	f1-score
gita	lbfgs	0.01	5000	0.66	0.614	0.86	0.46	0.717
gita	lbfgs	0.0001	2000	0.63	0.589	0.86	0.4	0.699
gita	lbfgs	0.0001	5000	0.62	0.579	0.88	0.36	0.698
gita	lbfgs	0.001	2000	0.6	0.571	0.8	0.4	0.667
fralusopark	lbfgs	0.0001	5000	0.586	0.586	0.773	0.369	0.667

Table 5.6: Independent experiment result using architecture 2.

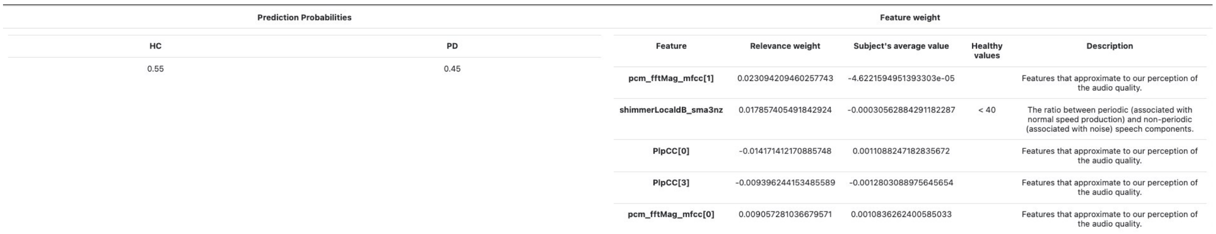
parameters to optimize, than architecture 2, which comprises a total of 52601 parameters. This difference makes architecture 1 much less resource-intensive, in both terms of time and computing power.

5.1.4 Model optimization

When comparing models' results per parameter, it is possible to find the best values for each parameter.
Smaller values for alpha (0.0001 and 0.001) consistently produced superior results when compared with 0.01. Considering language-dependent and semi language-dependent models, there is no clear difference between the use of the lbfgs and adam solvers. For both experiments, around half of the top five model parameterizations used each solver. In addition, for language-independent experiments, models using the lbfgs solver outperformed those using the adam solver. Between the top five model parameterizations of each architecture, only 1 was trained using adam (tables **5.5** and **5.6**). Lastly, comparing the results based on the number of maximum number of iterations ($\#interations$), there is no clear difference between models trained with $\#iterations = 2000$ and $\#iterations = 5000$ in any of the experiments performed. This shows that, in most cases, 2000 iterations should be sufficient to train the model, and convergence is reached without executing the maximum number of iterations.

5.2 Language Independency

Both architectures used during this work yielded an accuracy of 90% on the semi language-independent experiments. One the one hand, these results are inferior to the ones achieved on a similar work ((Orozco-Arroyave et al., 2016)), where the authors were able to achieve a maximum accuracy of 96% when training a model with a German dataset and 80% of a Spanish dataset and testing with the remaining 20%. On the other hand, this model was outperformed by architecture 1 when using the recall metric, producing recall values of 95%, whereas archi-



Prediction Probabilities

HC	PD
0.55	0.45

Feature weight

Feature	Relevance weight	Subject's average value	Healthy values	Description
pcm_fftMag_mfcc[1]	0.023094209460257743	-4.6221594951393303e-05		Features that approximate to our perception of the audio quality.
shimmerLocaldB_sma3nz	0.017857405491842924	-0.00030562884291182287	< 40	The ratio between periodic (associated with normal speed production) and non-periodic (associated with noise) speech components.
PlpCC[0]	-0.014171412170885748	0.0011088247182835672		Features that approximate to our perception of the audio quality.
PlpCC[3]	-0.009396244153485589	-0.0012803088975645654		Features that approximate to our perception of the audio quality.
pcm_fftMag_mfcc[0]	0.009057281036679571	0.0010836262400585033		Features that approximate to our perception of the audio quality.

Figure 5.1: Example of an explanation generated by LIME.

tecture 1 produced a recall of 100% for the top 3 model parameterizations. Contrary to this work, results produced by our model were inferior when using the specificity metric, where the authors were able to achieve a score of 97%, compared to the 75% produced by our model. Based on the recall metric, we can conclude that our solution has better ability to indicate when a subject belongs in the PD group. This contrasts with the ability to classify subjects from the HC group, where our model has an inferior performance. As previously described in section 5.1.3, architecture 1 produced an accuracy of 67% on the language-independent experiments. This result is slightly inferior to the one achieved on a different article (Orozco-Arroyave et al., 2016), where a language-independent model yielded an accuracy of 77% when trained with a Czech dataset and tested with a German dataset. Comparing the models using the recall and specificity metrics, the results are identical to the ones achieved on the semi language-independent models' comparison in this work. Our model produced a recall of 76% whereas the authors were only able to score 53% on this metric. On the other hand, architecture 1 produced a score of 58% on the specificity metric, significantly inferior to the 95% achieved by the other work.
It is possible to conclude that the performance of both architectures used in this work were not able to produce results at the state-of-the-art on the language independency topic. Regarding the recall metric, both architectures outperformed the state-of-the-art, which demonstrates better capacity in detecting PD.

5.3 Explainability

LIME was used to generate explanations for each test subject. These are local explanations, as they are able to explain the classification of each test subject. Results obtained following this process are described in section 5.3.1. By analyzing the complete set of explanations produced in this work, the global contribution (weight) of each feature was evaluated for the classification. Results for the global analysis are described in section 5.3.2.

5.3.1 Local Explanations

To generate an explanation, the top 5 features with the highest contribution to the diagnostic are selected. Figure 5.1 illustrates an explanation, containing the percentage attributed to each class (PD and HC), the features with the highest contribution to the diagnostic, their corresponding weights, the subject's average value on that feature, the range of normal values for a healthy subject (extracted from the bibliography), and a short description of the feature. This

feature	percentage of subjects	contribution (weight)
PlpCC[0]	14.783	0.057
PlpCC_de[1]	13.913	0.043
PlpCC_de_de[3]	13.261	0.044
pcm_fftMag_mfcc_de[1]	13.261	0.043
pcm_fftMag_mfcc[11]	12.391	0.042
pcm_fftMag_mfcc[8]	11.522	0.043
pcm_fftMag_mfcc[0]	11.304	0.058
pcm_fftMag_mfcc_de[0]	11.087	0.041
PlpCC_de[0]	11.087	0.043
pcm_fftMag_mfcc_de_de[2]	10.652	0.041

Table 5.7: Top 10 more common features on explanations.

feature	percentage of subjects	contribution (weight)
pcm_fftMag_mfcc[0]	11.30	0.058
PlpCC[0]	14.78	0.057
PlpCC[1]	6.09	0.051
pcm_fftMag_mfcc[12]	6.74	0.049
pcm_fftMag_mfcc[4]	8.04	0.049
PlpCC_de_de[0]	1.96	0.048
jitterLocal_sma3nz	3.26	0.047
pcm_fftMag_mfcc[2]	7.17	0.046
pcm_fftMag_mfcc_de[12]	5.87	0.046
pcm_fftMag_mfcc[6]	6.96	0.046

Table 5.8: Top 10 features ordered by contribution (weight) to explanations.

information provides a clearer insight of the model's classification to the medical professional. The percentage attributed to each class allows to evaluate the degree of confidence of the model in the decision, whereas the average features values can be compared to the normal range of values to check for abnormal parameters. Finally, the feature description links the mathematical definition of the features with its physical manifestation, thus simplifying the interpretation of the results by the medical professional.

5.3.2 Global Feature Contribution

The top 10 features were sorted by their frequency on the complete set of explanations produced in this work and by average contribution to the models' classification, (Tables 5.7 and 5.8).

PLP and MFCC are different mathematical representations of sound that simulate the way humans perceive it. These two sets of features constitute the majority of the top features with highest contribution to the largest number of test subjects (tables 5.7 and 5.8). Comparing the MFCCs (represented on the table as *pcm_fftMag_mfcc[n]*) and PLPs (represented on the table as *PlpCC[n]*), there are no significant differences between these features. In addition, jitter (*jitterLocal_sma3nz*) is also on the top features with higher weight contribution. Finally, F0, shimmer and HNR produce significant contributions to few test subjects (8.2% for F0, 7.4% for shimmer and 6.5% for HNR). These features' contributions are inferior to the ones shown on the table (0.0437 for F0, 0.0402 for HNR and 0.0379 for shimmer).

The global contribution (weight) for each feature can be observed in figure 5.2. The weight of the feature with highest global contribution is around 66% larger than the weight from the feature with lowest contribution. In addition, there is a significant difference between the weight of the top three features and the others, which can be defined as a threshold to separate the features into two groups (*relevant* and *irrelevant*).

The small difference between contribution (weight) values for consecutive features suggests a correlation between some features. The same trend applies when sorting features by percentage of subjects.

The best performing features are similar in both analysis, with a strong presence of MFCC and PLP group of features. A significant difference can be observed between the 5th and the 6th top features, which can also be defined as the threshold to separate the features into *relevant* and *irrelevant* groups.

Combining both analysis, the combined threshold can be defined as the top 5 features, meaning that this should be the group of features that the medical professional should focus on.

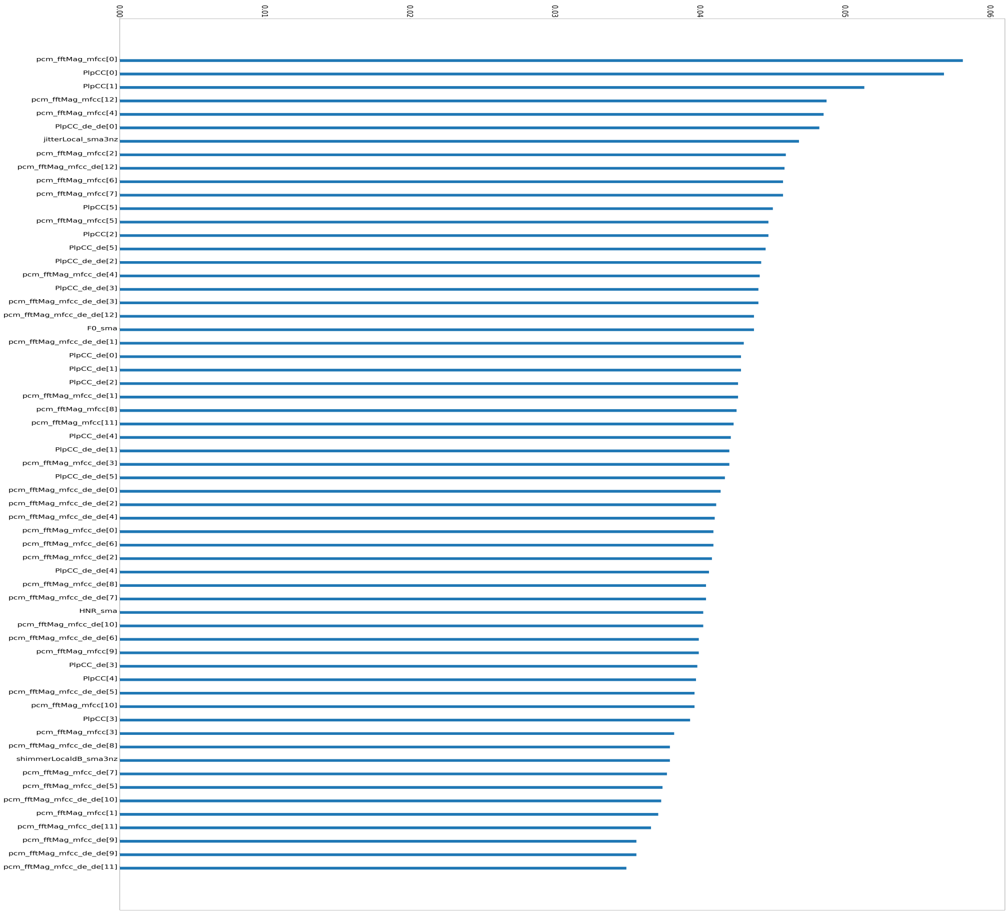


Figure 5.2: Global contribution (weight) by feature.

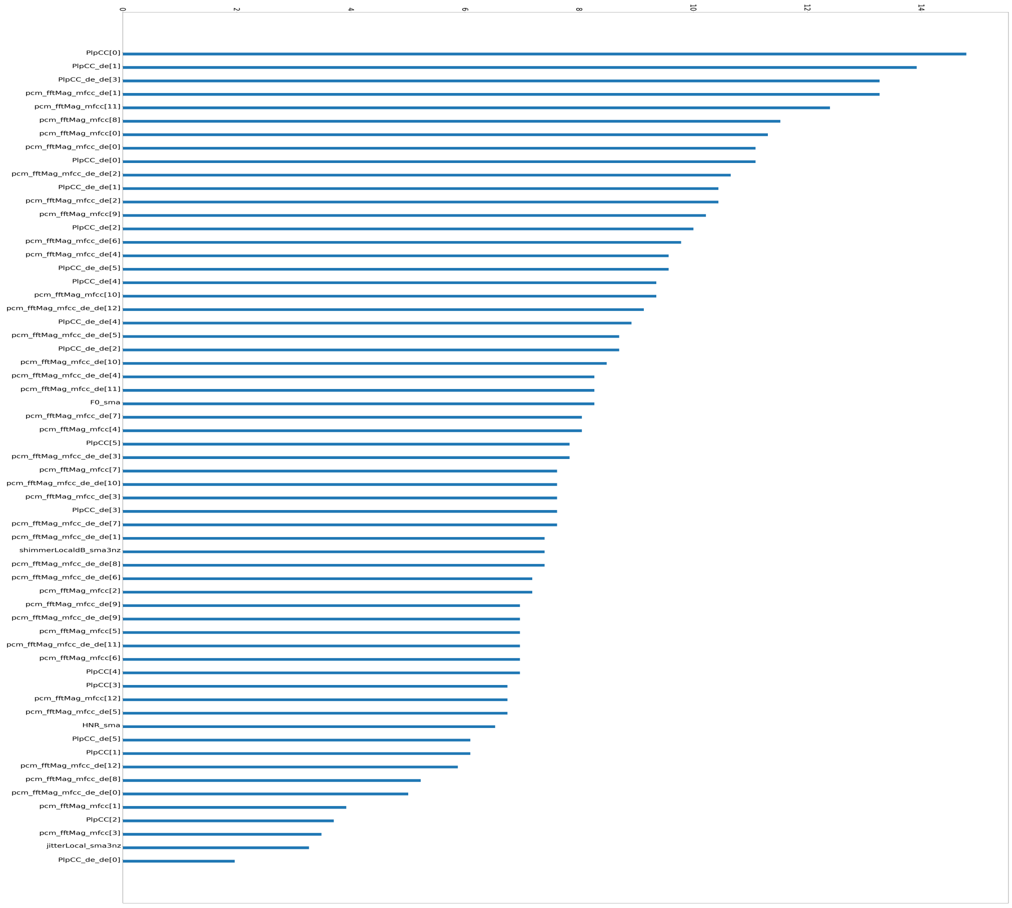


Figure 5.3: Percentage of subjects by feature.

6 Conclusions

There are a number of paths to continue this work.

First, the current pipeline presents some limitations that should be addressed. As previously described, there are complexity limitations associated with abstract features, such as PLPs and MFCCs. By using simpler features, such as Logarithmic Filter Banks (instead of MFCC) would increase the ability of the medical professional to understand, and therefore trust, the model's output. In addition, graphical representations of the physical manifestation of each feature can be added to the explanation. The normal values for some features, such as F0, depend on meta features (the normal values for F0 for males is between 105 and 160 Hz, while for females it's 175 to 245 *Hz*). Thus, adding the gender as a feature for the model would provide important information which could help improve the model's performance.

Both the classification and explanation pipeline's steps have further aspects to be explored. Namely, the similarity between the average contribution (weight) value of each feature on the explanation model suggested some correlation between features. This hypothesis can be further studied, using a model to evaluate the interactions between features, such as factorization machines. Detecting redundant features could help reduce the model's complexity, thus reducing resource necessity. Also, the results achieved on the semi language-independent experiments showed that there was no performance loss when training a model with two languages. Further analysis on the impact of varying the training percentage of the test language would shed light into the relation between data quantity used to re-train a model and the eventual performance loss. Moreover, this work focuses on explaining the diagnosis of each patient. A study on the global contributions of each feature could clarify the their individual importance to the PD classification task. This could be done by using LIME to generate global explanations', or by using models such as CAVs (Kim et al., 2018) or NAMs (Agarwal et al., 2020). Finally, both for the classification and explanations' steps, different models can be used to make a comparative analysis. This would allow to both assess the classification ability of multiple models, but also to compare the explanations generated by various models and the trust provided to the medical professionals.

The goal of generating explanations is to provide the medical professionals with a tool that can shed light into the *black-box* classification models. Thus, these models should be tested in real-world scenarios, to rate their adequacy to perform this task. During the real-world evaluation, a comparative analysis could be conducted between explainability models, in order to assess which ones provide more trust to the end users for the product (the medical professionals). This can be done by generating explanations for the same user using different explainability models and assessing the degree of confidence of the medical professional in each one of them. This evaluation could also lead to the conclusion that a combination of both methods provides more information, which would provide a higher level of trust by the medical professional on the classification models. Feature types (such as audio ou images) should also be compared, as to understand which are more adequate to be used by medical professionals. For example, the explanations generated by the model developed during this work could be compared with ones produced by the work described on section 3.3, in which LIME was used to explain PD diagnostic with SPECT DaTSCAN images of the brain.

Future work:
IMPROVE extraction/classification/explanation - Add meta features, such as gender, that condition the normal value of a feature (such as F0)
YES - Test simpler features, such as log-filter bank (instead of MFCC) and compare accuracies. **These would be easier to explain and to get normal/healthy values**
YES - Add images with physical ”representation” YES

EXTRA EXPERIMENTS - Nas experiências semi-independentes, reduzir a percentagem de treino do segundo dataset para encontrar o volume de dados necessário para re-adaptar um modelo numa língua para outra
YES - use factorization machines to generate compound explanations (combine multiple features) YES - Extend this work to generate global explanations at the same time
YES - test different models

REAL WORLD - Test the method in a real-world environment to assess the ability of the model to aid the diagnostic process
YES - During the test, other explainability methods, such as the ones presented on papers X and Y (LIME on brain). This will shed light to the method preferred by medical professionals, and may lead to the conclusion that a combination of both methods provides more information, which will provide a higher level of trust by the medical professional on the models.

Bibliography

Agarwal, R., Frosst, N., Zhang, X., Caruana, R., & Hinton, G. (2020). Neural additive models: Interpretable machine learning with neural nets.

Almeida, J., Filho, P., Carneiro, T., Wei, W., Damaševičius, R., Maskeliunas, R., & Albuquerque, V. (2019). Detecting parkinson's disease with sustained phonation and speech signals using machine learning techniques. *Pattern Recognition Letters*, *125*, 55-62. doi: https://doi.org/10.1016/j.patrec.2019.04.005

Andonie, R. (2010). Extreme data mining: Inference from small datasets. *International Journal of Computers, Communications & Contro*, *3*, 280-291.

Bach, S., Binder, A., Montavon, G., Müller, K., & Samek, W. (2016). Analyzing classifiers: Fisher vectors and deep neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2912-2920.

Bakar, Z., Tahir, N., & Yassin, I. (2010). Classification of parkinson's disease based on multilayer perceptrons neural network. In *2010 6th international colloquium on signal processing its applications* (p. 1-4). doi: 10.1109/CSPA.2010.5545301

Bang, Y., Min, K., Sohn, Y., & Cho, S. (2013). Acoustic characteristics of vowel sounds in patients with parkinson disease. *NeuroRehabilitation*, *32*, 649–654.

Bot, B., Suver, C., Neto, E., Kellen, M., Klein, A., Bare, C., ... Trister, A. (2016). The mpower study, parkinson disease mobile data collected using researchkit. *Scientific Data*, *3*.

Braga, D., Madureira, A., Coelho, L., & Ajith, R. (2019). Automatic detection of parkinson's disease based on acoustic analysis of speech. *Engineering Applications of Artificial Intelligence*, *77*, 148-158. doi: https://doi.org/10.1016/j.engappai.2018.09.018

Chen, I., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2020). Ethical machine learning in health care. *The Hastings Center Report*. doi: https://doi.org/10.1146/

Choi, B., Shim, G., Jeong, B., & Jo, S. (2020). Data-driven analysis using multiple self-report questionnaires to identify college students at high risk of depressive disorder. *Scientific Reports*, *10*. doi: 10.1038/s41598-020-64709-7

Chung, J., Nagrani, A., & Zisserman, A. (2017). Voxceleb: A large-scale speaker identification dataset. In *Interspeech* (p. 2616–2620).

Chung, J., Nagrani, A., & Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. In *Interspeech* (p. 2616–2620).

Clinic, M. (2020). *Parkinson's disease.* https://www.mayoclinic.org/diseases-conditions/parkinsons-disease/symptoms-causes/syc-20376055. (Online; accessed 23 December 2020)

Das, A., & Rad, P. (2020). *Opportunities and challenges in explainable artificial intelligence (xai): A survey.*

Despotovic, V., Skovranek, T., & Schommer, C. (2020). Speech based estimation of parkinson's disease using gaussian processes and automatic relevance determination. *Neurocomputing*, *401*, 173-181. doi: https://doi.org/10.1016/j.neucom.2020.03.058

Dorsey, E., Sherer, T., Okun, M., & Bloem, B. (2018, 12). The emerging evidence of the parkinson pandemic. *Journal of Parkinson's Disease*, *8*, S3-S8. doi: 10.3233/JPD-181474

Erdogdu Sakar, B., Isenkul, M., Sakar, C., Sertbas, A., Gurgen, F., Delil, S., ... Kursun, O. (2013). *Collection and analysis of a parkinson speech dataset with multiple types of sound recordings* (Vol. 17).

Eyben, F., Wöllmer, M., & Schuller, B. (2010). opensmile – the munich versatile and fast open-source audio feature extractor. In (p. 1459-1462). doi: 10.1145/1873951.1874246

Fahn, S., & Elton, R. (1987). Unified parkinson's disease rating scale. *Recent Development in Parkinson's Disease*, *2*.

Fish, J. (2011). Unified parkinson's disease rating scale. In *Encyclopedia of clinical neuropsychology* (pp. 2576–2577).

Fleming, R., Zeisel, J., & Bennett, K. (2020). World alzheimer report 2020: Design dignity dementia: dementia-related design and the built environment, volume 1.

Froelich, W., Wróbel, K., & Porwik, P. (2015). Diagnosis of parkinson's disease using speech samples and threshold-based classification. *Journal of Medical Imaging and Health Informatics*, *5*, 1358-1363. doi: 10.1166/jmihi.2015.1539

Goberman, A., & Elmer, L. (2005). Acoustic analysis of clear versus conversational speech in individuals with parkinson disease. *Journal of Communication Disorders*, *38*, 215-230.

Godino-Llorente, J., Shattuck-Hufnage, S., Choi, J., Moro-Velázquez, L., & Gómez-García, J. (2017). Towards the identification of idiopathic parkinson's disease from the speech. new articulatory kinetic biomarkers.

Goetz, C., Poewe, W., Rascol, O., Sampaio, C., Stebbins, G., Fahn, S., ... Van Hilten, B. (2003). The unified parkinson's disease rating scale (updrs): Status and recommendations. *Movement Disorder*, *18(7)*.

Goetz, C., Poewe, W., Rascol, O., Sampaio, C., Stebbins, G., Fahn, S., ... Tilley, B. (2008). Mds-updrs: The mds-sponsored revision of the unified parkinson's disease rating scale..

Harel, B., Cannizzaro, M., Cohen, H., Reilly, N., & Snyder, P. (2004). Acoustic characteristics of parkinsonian speech: a potential biomarker of early disease progression and treatment. *Journal of Neurolinguistics*, *17*, 439–453.

Harrell, L., Marson, D., Chatterjee, A., & Parrish, J. (2000). The severe mini-mental state examination: A new neuropsychologic instrument for the bedside assessment of severely impaired patients with alzheimer disease. *Alzheimer Disease and Associated Disorders*, *14*. doi: https://doi.org/10.1097/00002093-200007000-00008

Health Direct. (2019). *Mini-mental state examination (mmse).* https://www.healthdirect.gov.au/mini-mental-state-examination-mmse. (Online; accessed 23 December 2020)

Hely, M., Reid, W., Adena, M., Halliday, G., & Morris, J. (2008). The sydney multicenter study of parkinson's disease: the inevitability of dementia at 20 years. *Movement Disorders Journal*, *23*, 837–844.

Hermansky, H., Morgan, N., Bayya, A., & Kohn, P. (1992). Rasta-plp speech analysis technique. In *Icassp-92: 1992 ieee international conference on acoustics, speech, and signal processing* (Vol. 1, p. 121-124).

Ho, A., Iansek, R., Marigliani, C., Bradshaw, J., & Gates, S. (1998). Speech impairment in a large sample of patients with parkinson's disease. *Behavioural neurology*, *11*, 131–137.

Hoehn, M. M., & Yahr, M. D. (1967). Parkinsonism. *Neurology*, *17*(5), 427–427. Retrieved from https://n.neurology.org/content/17/5/427 doi: 10.1212/WNL.17.5.427

Holzinger, A., Biemann, C., Pattichis, C., & Kell, D. (2017). What do we need to build explainable ai systems for the medical domain?

Jaeger, H., Trivedi, D., & Stadtschnitzer, M. (2019). *Mobile device voice recordings at king's college london (mdvr-kcl) from both early and advanced parkinson's disease patients and healthy controls [data set].* doi: http://doi.org/10.5281/zenodo.2867216

Kent, T., Vorperian, H., Kent, J., & Duffy, J. (2003). Voice dysfunction in dysarthria: application of the multi-dimensional voice program. *Journal of Communication Disorders*, *36*, 281–306.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). *Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav).*

Lapuschkin, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, *10*. doi: 10.1371/journal.pone.0130140

LeCun, Y., & Cortes, C. (2010). MNIST handwritten digit database.

Little, M., McSharry, P., Costello, D., & Moroz, I. (2007). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMedical Engineering OnLine*.

London, A. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *The Hastings Center Report*, *49*, 15-21. doi: 10.1002/hast.973

Lopez-de Ipiña, K., Alonso, J., Travieso, C., Solé-Casals, J., Ezeiza, A., Faundez-Zanuy, M., ... Calvo, P. (2014). Feature selection for automatic analysis of emotional response based on nonlinear speech modeling suitable for diagnosis of alzheimer's disease. *Neurocomputing*, *150*. doi: 10.1016/j.neucom.2014.05.083

Lopez-de Ipiña, K., Solé-Casals, J., Eguiraun Martinez, H., Alonso, J., Travieso, C., Ecay, M., ... Beitia, B. (2014). Feature selection for spontaneous speech analysis to aid in alzheimer's disease diagnosis: A fractal dimension approach. *Computer Speech & Language*, *30*. doi: 10.1016/j.csl.2014.08.002

Magesh, P., Myloth, R., & Rijo, T. (2020). An explainable machine learning model for early detection of parkinson's disease using lime on datscan imagery. *Computers in Biology and Medicine*, *126*. doi: https://doi.org/10.1016/j.compbiomed.2020.104041

Mayo Clinic. (2019). *Dementia.* https://www.mayoclinic.org/diseases-conditions/dementia/symptoms-causes/syc-20352013. (Online; accessed 23 Setember 2020)

Moreno, A., Poch-Olivé, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J., & Nadeu, C. (1993). Albayzin speech database: design of the phonetic corpus. In *Eurospeech.*

Moro-Velazquez, L., Gomez-Garcia, J., Godino-Llorente, J., & Dehak, N. (2019). A forced gaussians based methodology for the differential evaluation of parkinson's disease by means of speech processing. *Biomedical Signal Processing and Control*, *48*, 205–220.

Moro-Velázquez, L., Gomez-Garcia, J., Godino-Llorente, J., Grandas-Perez, F., Shattuck-Hufnagel, S., Yagüe-Jimenez, V., & Dehak, N. (2019). Phonetic relevance and phonemic grouping of speech in the automatic detection of parkinson's disease. *Scientific Reports*, *9*.

Moro-Velázquez, L., Gómez-García, J., Godino-Llorente, J., Villalba, J., Orozco-Arroyave, J., & Dehak, N. (2018). Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect parkinson's disease. *Applied Soft Computing*, *62*, 649–666.

Moro-Velázquez, L., Villalba, J., & Dehak, N. (2020). Using x-vectors to automatically detect parkinson's disease from speech. In *Icassp 2020 - 2020 ieee international conference on acoustics, speech and signal processing* (p. 1155-1159).

Naranjo, L., Pérez, C., Campos-Roca, Y., & Martín, J. (2016). Addressing voice recording replications for parkinson's disease detection. *Expert Systems With Applications*, *46*. doi: http://dx.doi.org/10.1016/j.eswa.2015.10.034

National Institute of Aging. (2017). *Parkinson's disease.* https://www.nia.nih.gov/health/parkinsons-disease. (Online; accessed 23 December 2020)

Orozco-Arroyave, J., Arias-Londoño, J., Vargas-Bonilla, J., Gonzalez-Rátiva, M., & Nöth, E. (2014). New spanish speech corpus database for the analysis of people suffering from parkinson's disease. In *Lrec.*

Orozco-Arroyave, J., Honig, F., Arias-Londoño, J., Vargas-Bonilla, J. F., Daqrouq, K., Skodda, S., … Noth, E. (2016). Automatic detection of parkinson's disease in running speech spoken in three different languages. *Journal of the Acoustical Society of America*, *139*, 481–500.

Pagan, F. (2012). Improving outcomes through early diagnosis of parkinson's disease. *American Journal of Managed Care*, *18*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Perez, K., Ramig, L., Smith, M., & Dromey, C. (1996). The parkinson larynx: tremor and videostroboscopic findings. *Journal of voice : official journal of the Voice Foundation*.

Petsiuk, V., Das, A., & Saenko, K. (2018). Rise: Randomized input sampling for explanation of black-box models. *CoRR*.

Pfob, A., Sidey-Gibbons, C., Lee, H.-B., Tasoulis, M. K., Koelbel, V., Golatta, M., ... Kuerer, H. (2020). Identification of breast cancer patients with pathologic complete response in the breast after neoadjuvant systemic treatment by an intelligent vacuum-assisted biopsy. *European Journal of Cancer*, *143*, 134-146. doi: 10.1016/j.ejca.2020.11.006

Pinto, S., Cardoso, R., Sadat, J., Guimarães, I., Mercier, C., Santos, H., ... Ferreira, J. (2016). Dysarthria in individuals with parkinson's disease: A protocol for a binational, cross-sectional, case-controlled study in french and european portuguese (fralusopark). *BMJ Open*, *6*. doi: 10.1136/bmjopen-2016-012885

Pompilli, A., Abad, A., Romano, P., Pavão Martins, I., Cardoso, R., Santos, H., ... Ferreira, J. (2017). Automatic detection of parkinson's disease: An experimental analysis of common speech production tasks used for diagnosis. *Text, Speech, and Dialogue. TSD 2017. Lecture Notes in Computer Science*, *10415*. doi: https://doi.org/10.1007/978-3-319-64206-2_46

Proença, J., Perdigão, F., Veira, A., Candeias, S., Lemos, J., & Januário, C. (2014). Characterizing parkinson's disease speech by acoustic and phonetic features. *Computational Processing of the Portuguese Language*, 24-35.

Punn, N., & Agarwal, S. (2020). Automated diagnosis of covid-19 with limited posteroanterior chest x-ray images using fine-tuned deep neural networks. *Applied Intelligence*, 1-14. doi: 10.1007/s10489-020-01900-3

Rai, M., Yadav, A., Ingle, A. P., Reshetilov, A., Blanco-Prieto, M. J., & Feitosa, C. (2019). Neurodegenerative diseases: The real problem and nanobiotechnological solutions. In *Nanobiotechnology in neurodegenerative diseases* (pp. 1–17). Springer International Publishing.

Ribeiro, M., Singh, S., & Guestrin, C. (2016). "why should i trust you?" explaining the predictions of any classifier. In *Kdd '16: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (p. 1135–1144).

Rusz, J., Cmejla, R., Tykalová, T., Ruzickova, H., Klempír, J., Majerova, V., ... Růika, E. (2013). Imprecise vowel articulation as a potential early marker of parkinson's disease: effect of speaking task. *The Journal of the Acoustical Society of America*, 2171-2181.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *CoRR*.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *Icml*.

Tombaugh, T., & McIntyre, N. (1992). The mini-mental state examination: a comprehensive review. *Journal of the American Geriatrics Society*, *40*. doi: https://doi.org/10.1111/j.1532-5415.1992.tb01992.x

Tracy, J., Özkancab, Y., Atkins, D., & Ghomi, R. (2020). Investigating voice as a biomarker: Deep phenotyping methods for earlydetection of parkinson's disease. *Journal of Biomedical Informatics*, *104*. doi: https://doi.org/10.1016/j.jbi.2019.103362

Tsanas, A., Little, M., McSharry, P., & Ramig, L. (2009). Accurate telemonitoring of parkinson's disease progression by non-invasive speech tests. *IEEE Transactions on Biomedical Engineering*.

Tysnes, O.-B., & Storstein, A. (2017). Epidemiology of parkinson's disease. *Journal of neural transmission*, *124*, 901–905.

Vellido, A. (2019). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*. doi: https://doi.org/10.1007/s00521-019-04051-w

Wan, S., Liang, Y., Zhang, Y., & Guizani, M. (2018). Deep multi-layer perceptron classifier for behavior analysis to estimate parkinson's disease severity using smartphones. *IEEE Access*. doi: 10.1109/ACCESS.2018.2851382

Williams-Gray, C., Mason, S., Evans, J., Foltynie, T., Brayne, C., Robbins, T., & Barker, R. (2013). The campaign study of parkinson's disease: 10-year outlook in an incident population-based cohort. *Journal of Neurology, Neurosurgery and Psychiatry*, *84*, 1258–1264.

World Health Organization. (2020). *Dementia.* https://www.who.int/news-room/fact-sheets/detail/dementia. (Online; accessed 23 Setember 2020)

Yaman, O., Ertam, F., & Tuncer, T. (2020). Automated parkinson's disease recognition based on statistical pooling method using acoustic features. *Medical Hypotheses*, *135*. doi: https://doi.org/10.1016/j.mehy.2019.109483

I Appendices

Glossary

AD Alzheimer's Disease. 1, 7
AI Artificial Intelligence. 1, 9
ARD Automatic Relevance Determination. 6, 7

BC Boosting Classifier. 7

CAV Concept Activation Vectors. 11

DDK Diadochokinesia. 5, 6, 15, 16
DeepLIFT DeepLIFT. 12
DL Deep Learning. 9
DMLP Deep Multi-Layer Perceptrons. 8
DNN Deep Neural Networks. 8, 10, 12
DT Decision Trees. 7

F0 Fundamental Frequency. 3, 4, 6, 7, 17, 23

GMM-UBM Gaussian Mixture Model - Universal Background Model. 5–8
GNR Glottal-to-Noise Ratio. 4, 6
GPC Gaussian Process Classification. 5–7

H&Y Hoehn and Yahr. 4
HC Healthy Controls. 3, 5–7, 16–19, 21, 22
HNR Harmonics-to-Noise Ratio. 4, 6, 17, 23

IG integrated Gradients. 10

KNN K-Nearest Neighbours. 7, 8

LIME Local Interpretable Model-agnostic Explanation. 9, 12
LR Logistic Regression. 7, 8
LRP Layer-wise Relevance Propagation. 9, 12

MDS Movement Disorder Society. 4
MFCC Mel-frequency cepstral coefficients. 6, 7, 23
ML Machine Learning. 1, 4–9, 11, 13
MLP Multi-Layer Perceptrons. 7, 17, 19, 21

NAM Neural Additive Models. 11
NHR Noise-to-Harmonics Ratio. 4, 6
NN Neural Networks. 6, 7, 11, 12

PD Parkinson's Disease. i, 1–9, 13, 15–19, 21, 22

PLDA Probabilistic Linear Discriminant Analysis. 6

PLP Perceptual Linear Predictive. 17, 23

Rasta-PLP Rasta-Perceptual Linear Predictive. 6

RF Random Forests. 6–8

RISE Randomized Input Sampling for Explanation. 10, 12

SVM Support Vector Machines. 6, 7

TDU Text-dependent Utterances. 5, 6

UPDRS Unified Parkinson's Disease Rating Scale. 4

VTI Voice Turbulence Index. 4, 6

XAI Explainable Artificial Intelligence. 2, 9

	Male	Female
F0 (Hz)	105-160	175-245
Jitter (%)	< 1.04	
Shimmer (%)	< 3.81	
HNR (dB)	< 20 (/a/, /i/), < 40 (/u/)	

Table 1: Feature values for healthy subjects.

	Description
F0	The number of open/close cycles of the glottis.
Jitter	Measures the frequency variation between cycles. Affected by lack of control on the vocal cords vibration.
Shimmer	Measures the amplitude variation between cycles.
HNR	The ratio between periodic (associated with normal speed production) and non-periodic (associated with noise) speech components.
MFCC	Features that approximate to our perception of the audio quality
PLP	Features that approximate to our perception of the audio quality

Table 2: Acoustic features description.