Explaining Parkinson's Disease Computational Diagnostic based on Speech Analysis

Artur Oliveira Fortunato

Instituto Superior Técnico, Universidade de Lisboa
Lisboa, Portugal
artur.fortunato@tecnico.ulisboa.pt

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 biomarker for the detection of PD. The present work had two objectives. First, we aimed at assessing the performance of a language-independent model for the Parkinson's Disease (PD) diagnostic task. For this work, three datasets from different languages were used. A baseline approach (a model trained and tested with speech from the same language) achieved a maximum accuracy of 90%. An intermediate step was also taken, where a model was trained with speech from one language and part of the speech contained in a different dataset (in a different language) and tested with the remaining part of the speech from the second dataset. This semi language-independent model's performance was similar to the baseline's performance. These results demonstrated the ability of our model to be re-trained with new data from a new language and be extended to patients speaking the new language. Next, the language-independent model was trained, reaching a maximum accuracy of 67% and a recall value of 76%. Although the accuracy of our model is lower than the state-of-the-art (77%), the recall, which represents the capacity to detect PD patients, is far better than the best previous work (53%). Second, the LIME explainability model was used to generate an explanation report for each diagnostic produced by the classification model. The report includes the probability of a subject belonging to each class (PD or Healthy Controls (HC)) and the top five features with the highest contribution to the model's classification. Each feature includes the average value, the range of values of a healthy individual, its contribution weight to the classification, and a small description. This information helps the clinician to understand the computational diagnostic, thus providing enhanced trust in the model. An evaluation of the global contribution of each feature concluded that both MFCC and PLP features provide the models with more relevant information than Fundamental Frequency (F0), Harmonics-to-Noise Ratio (HNR), jitter, and shimmer. This work contributed to increase the usefulness of machine learning models for the automatic PD detection. Its contribution was a PD detection approach that can be extended to any language. Furthermore, the explainability model applied herein facilitates the understanding and fosters the adoption of PD diagnostic computational models in real medical practice.

Keywords: Machine Learning, Speech, Explainability, Interpretability.

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) [53].

PD is the second most common neurodegenerative disease. It was estimated that 1% of people over 60 years old are affected with PD [60]. 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 [16]. 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 [63], reaching over 80% after 20 years [24].

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 [43].

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. A common problem in these experiments is the lack of training data. A solution for this problem is to train a global model that could be used for patients with different characteristics than the ones from the training data (for example, for speech-based models, it would be able to make diagnostics using speech in a lan-

guage different from the one used to train the model). In addition, 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 [33]. Replacing medical decision-making with non-explainable, black-box ML models, can be contravening with the profound ethical responsibilities of clinicians [31]. Consequently, the lack of explainability and interpretability of ML models used in these areas can seriously limit their chances of adoption in real practice [61]. 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 [27].

This work's objective was twofold. First, we trained a model that was able to detect PD from speech in a language different from the training data. Second, we used an explainability model to generate human-understandable explanations for the given classification (Parkinson or Healthy) of each patient 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 6 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 [40].

2.1 Symptoms

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) [13].

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) [23], pause duration [23] or vowel space time [21] have been shown to distinguish PD from Healthy Controls (HC). 90% of PD patients are reported to have speech and voice disorders [20], 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 [46], 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 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.

2.3.1 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 [22], 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), and text reading.

Several speech production tasks to detect PD were tested [50] – 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 [49], 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 [30] and the Parkinson's Telemonitoring Dataset [59]. 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% [15].

To study the relevance of each phonemic group in detecting PD, three datasets were used – GITA [41], Neurovoz [36], and CzechPD [56]. 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 [35], the model yielded an classification accuracy of 94% for the CzechPD dataset, 89% for Neurovoz, and 84% for GITA [37].

Sustained vowels and text reading tasks were tested to differentiate PD from HC [8]. The authors use three datasets – Proença [51] (containing audio from 22 PD patients in European Portuguese), UCI [17] (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.

2.3.2 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 [46], 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 [6] and shimmer [29].

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) [15]. 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 [37]. 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) [25] 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 [11] and 2 [12] 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/*) [38].

2.3.3 Classification models Most of the available datasets for this task are very small, considering the usual size for a classification problem. This characteristic 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 [2]. 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 [62].

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% [50].

A set of classifiers was used on two PD datasets [15]. 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%. 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 [39] 240 recordings were retrieved [64]. 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 [64]. 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 [7] (2268 corresponding to Audio/Visual Emotion and Depression Recognition Challenge (AVEC) 2013 and 62 corresponding to GeMAPS) [58]. 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% [37].

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 [32]. The Levenberg-Marquardt and Scaled Conjugate Gradient methods were tested as training algorithms for an MLP [5]. 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 [17], a set of 23 features was extracted for PD classification [62]. 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.

2.3.4 Language independency As one of the goals 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 [42]. 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% [37].

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 [14].

Explaining the decisions made by a black-box model requires knowledge of its internal operations [14], 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 [57]. 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 [55]. 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 [4]. 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 [14]. 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 [33]). Some models, such as Local Interpretable Model-agnostic Explanation (LIME) [55], base their explanations on activations or parameters of the black-box models, using simpler surrogate models [14].

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 [14]. 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 [1].

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) [47].

Consider $f : I \to \mathbb{R}$ 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 ($I = \{I | I : \Lambda \to \mathbb{R}\}$). 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 $M : \Lambda \to \{0, 1\}$ 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,

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

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 [14].

In 2020, the concept of Neural Additive Models (NAM) was proposed [1]. The explanations are created by shape functions, relative to each input feature. To parameterize 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:

$$\mathcal{L}(\Theta) = E_{x,y \sim \mathcal{D}}[l(x, y; \Theta) + \lambda_1 \eta(x; \theta)] + \lambda_2 \gamma(\Theta) \quad (2)$$

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

$$l(x, y; \Theta) = (\beta^{\Theta} + \sum_{k=1}^{K} f_k^{\Theta}(x_k) - y)^2 \quad (3)$$

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.

LIME is an algorithm that uses *local interpretable representations* of the classification data to generate an output that can be interpreted by humans [55]. 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:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad (4)$$

The algorithm creates sample instances instances 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. For example, considering an input $x = [1, 2, 3, 4, 5]$ and a mask $x' = [1, 1, 1, 1, 0]$, z' could be [1,0,1,1,0] (ignoring the second value of the input). Thus, z can be defined as $z = z' \odot x = [1, 0, 3, 4, 0]$. Considering $\mathcal{Z}$ as the set of all perturbed z' with the label $f(z)$, equation 4 is used to calculate the explanation.

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 [27], 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 [9].

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 [33]. 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 [48], identification of individuals with high-risk of depressive disorder [10], and early detection of COVID-19 [52]. 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 1 shows the pipeline for the system's architecture.

4.1 Corpus Description

Most datasets available for this task have insufficient data to train neural models [2]. Nevertheless, few common speech production tasks are available in the datasets. As some datasets contain speech from PD patients and share commons speech production tasks, they can be combined to produce sufficiently long collections of data that can be used for neural models [8], [15], [37], [38]. Different datasets were used for training and testing, or to combine instances from different datasets in the training and/or testing sets [42], all proving to be accurate in the PD classification task.

This study used 3 datasets for training and testing the model – FraLusoPark [49], GITA [41], and Mobile Device Voice Recordings at King's College London (MDVR_KCL) [28].

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 the following set of speech production tasks: sustain the vowel *a* at a steady pitch, hold the vowel *a* during their maximum phonation time 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, and free conversation for 3 minutes. In the scope of the present study, we only consider the Portuguese speakers of this dataset, as the audios from the french patients could not be accessed.

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.

Lastly, the MDVR_KCL dataset was recorded in the context of phone calls, in an acoustically-controlled environment. The dataset contains speech 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 [26]. The participants were asked to read a text ("The north wind and the sun" or "Tech. Engin. Computer applications in geography snippet"). Additionally, the inter-

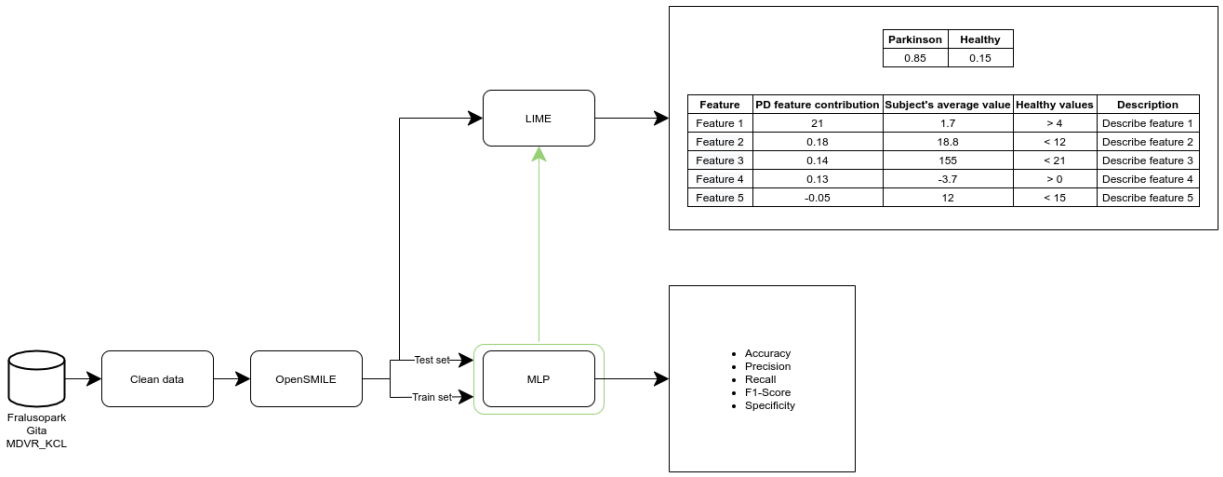

Fig. 1: Pipeline of the developed model.

viewer 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 the interviews (including silences and speech segments from the interviewers) 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 deleted. Finally, audio segments containing speech from interviewers were also eliminated.
After data processing, the datasets were reduced to one file per subject. Total duration and remaining metadata on each dataset is presented in table 1.

4.3 Feature Extraction

In order to extract the features, the openSMILE [18] 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, *prosodyAcf2.conf* for prosody features (F0 and HNR), and *GeMAPSv01b.conf* to extract jitter and shimmer.
OpenSMILE was configured to use a sliding window of $25ms$ with a frame step of $10ms$. After the extraction, each participant was represented by a sequence of feature vectors, 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 dataset. 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 (90%), 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 [45] 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 $\mathbf{N}$ neurons (where $\mathbf{N}$ is the number of input features), a fully-connected hidden layer with $\mathbf{N+1}$ neurons and an output layer with 1 neuron, whose value represents the probability of the test subject to be classified as PD [32]. The second architecture also contains an input layer with $\mathbf{N}$ 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

Dataset	Total time	PD/HC (%)	M/F (%)	Average age
FraLusoPark	113m43s	54/46	51/49	65.1
Gita	30m8s	50/50	50/50	61.7
MDVR_KCL	65m3s	42/58	38/62	n/a

Table 1: Information on the used datasets.

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^{-4}, 10^{-3}, 10^{-2}), maximum number of iterations (1000, 2000, 5000) and solver for weight optimization (lbfgs, sgd, adam).

4.5 Explanation generation

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.
The selected model was LIME [55]. 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 2 and **??** show the complete list of normal values and descriptions for each feature.
Additionally, a global evaluation of each feature's relevance was conducted. In order to conduct this analysis, the top five classification model configurations for each experiment for each MLP were used. An explanation was generated for each patient from the testing set, and the result contribution of each feature was averaged. Additionally, the percentage of patients for each of the top five features with highest contribution was calculated. Thus, we shed light into the relative quality of each feature in distinguishing PD patients from healthy subjects.

4.6 Model Evaluation

To evaluate the classification model's performance, multiple metrics have been selected, namely*accuracy* (which allows to evaluate the % of subjects correctly diagnosed), *precision* (that yields the fraction of subjects diagnosed with PD that were correctly classified, *recall* (which quantifies the percentage of PD subjects that were correctly diagnosed), *F1-score* that allows to evaluate *precision* and *recall* in the same metric, and *Specificity* which measures the fraction of subjects classified as HC that were correctly diagnosed.

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 results, average values of each feature were obtained from the bibliography (see table 2) and are shown along with the values in each explanation, in order to compare each subject's feature values with its 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. Furthermore, a comparative analysis between the global contribution of each acoustic feature was performed, therefore allowing to compare feature's relevance for the model's diagnostic.

5 Results and Discussion

This chapter presents the results and discussion. First, classification experiments are presented, Results and discussion 5 describing all three experiment types (baseline, semi language-independent, and language-independent) and discussing model optimization. Next, the language independency topic is discussed, followed by the explainability model's results and a comparative analysis on the relevance of each acoustic feature.

5.1 Classification Experiments

In this work, three types of experiments were conducted, each using two different MLP architectures, as described in the previous chapter. Results are shown in tables 3 and 4 (for the baseline experiments), 5 and 6 (for the semi language-independent experiments), and 7 and 8 (for the language-independent experiments). These tables show the five MLP parameter parameterizations with higher accuracy for each experiment. Tables 3, 5 and 7 present the results for architecture 1, whereas tables 4, 6 and 8 show the results for architecture 2.

	Male	Female	Reference
F0 (Hz)	105-160	175-245	[44]
Jitter (%)		< 1.04	[3]
Shimmer (%)		< 3.81	[3]
HNR (dB)	< 20 (/a/, /i/), < 40 (/u/)		[19]

Table 2: Feature values for healthy subjects.

5.1.1 Baseline experiments Both architectures 1 and 2 of the MLP yielded an accuracy of 90% with the best parameterization (tables 3 and 4 and figure 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.
Initial experiments using either the *sgd* solver or $\#iterations = 1000$ produced significantly lower results compared to the other values. Therefore, these two values were removed. 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 value smaller values (0.0001 and 0.001) for the *alpha* parameter, comparing to the results obtained using larger values (0.01). 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 62 bias. A larger number of parameters requires more iterations for the model to converge.
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. Architectures 1 and 2 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 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 5 and 6 and figure 3). 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 a similar work that tested semi language-independent models [42], which suggests that these models can be 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 7 and 8 and figure 4).

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 3: 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 4: Baseline experiment result using architecture 2.

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

Table 5: Semi language-independent experiment result using architecture 1. Dataset column legend: **M** - MDVR_KCL, **F** - FralusoPark, **G** - Gita. First dataset was used entirely for training, the second one was partially used for training and partially for testing.

dataset	solver	alpha	max. iterations	accuracy	precision	recall	specificity	f1-score
M + G	lbfgs	0.001	5000	0.9	1.0	0.8	1.0	0.889
F + G	lbfgs	0.0001	2000	0.8	0.75	0.75	0.833	0.75
M + G	adam	0.0001	5000	0.8	1.0	0.667	1.0	0.8
M + G	adam	0.001	5000	0.8	0.6	1.0	0.714	0.75
M + G	lbfgs	0.0001	5000	0.8	0.6	1.0	0.714	0.75

Table 6: Semi language-independent experiment result using architecture 2. Dataset column legend: **M** - MDVR_KCL, **F** - FralusoPark, **G** - Gita. First dataset was used entirely for training, the second one was partially used for training and partially for testing.

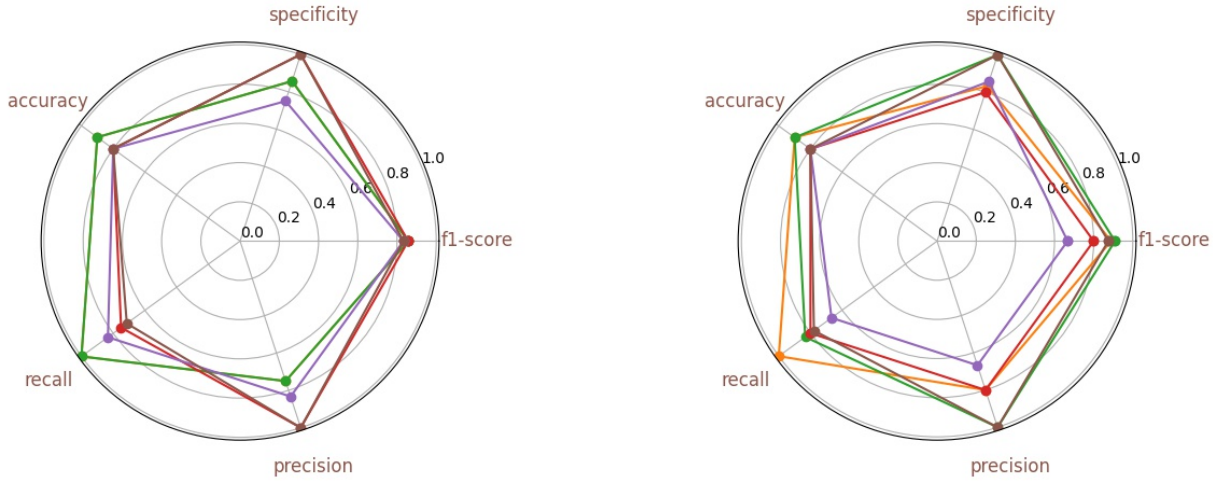

Fig. 2: Baseline experiment result using architecture 1 (left) and architecture 2 (right). Correspondence to table 3 and 4: row 1 is orange, row 2 is green, row 3 is red, row 4 is purple, row 5 is brown.

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 baseline and semi language-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 $alpha = 0.01$. 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 3906 parameters to optimize, than architecture 2, which comprises a total of 52801 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 7 and 8). 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 ([42]), 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

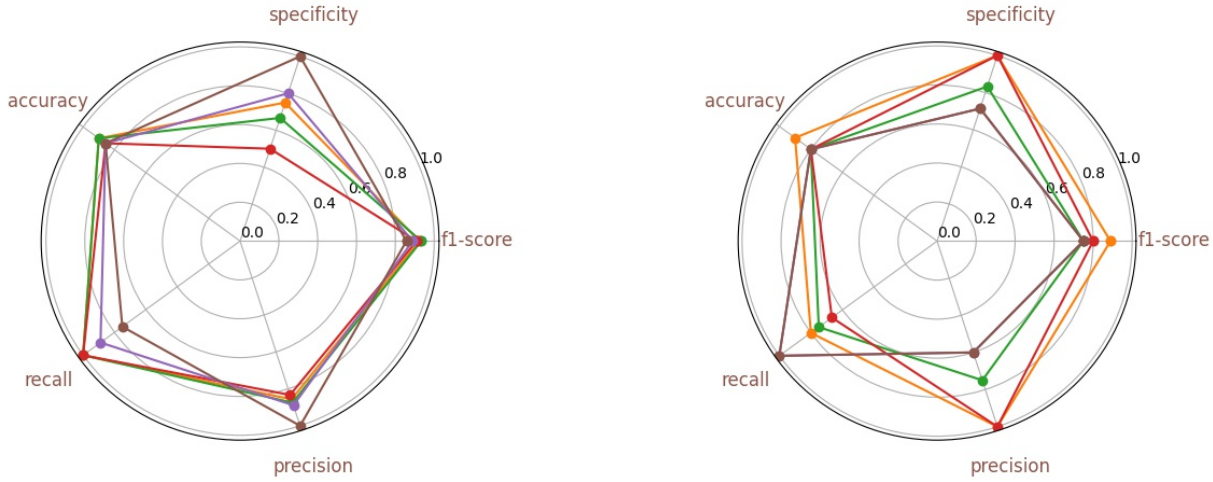

Fig. 3: Semi language-independent experiment result using architecture 1 (left) and architecture 2 (right). Correspondence to table 5 and 6: row 1 is orange, row 2 is green, row 3 is red, row 4 is purple, row 5 is brown.

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 7: Independent experiment result using architecture 1.

1 when using the recall metric, producing recall values of 95%, whereas architecture 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 [42], 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 with highest accuracy 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 state-of-the-art results 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 five features with the highest contribution to the diagnostic were selected. Figure 5 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 (values ranging between [-1,1]), 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 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 value can be compared to the normal range of values to

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 8: Independent experiment result using architecture 2.

feature	percentage of subjects	contribution (weight)
PLP[0]	77.2	5.4
MFCC[0]	65.3	4.8
PLP[1]	55.4	3.9
MFCC[1]	44.6	4.7
MFCC[12]	38.6	5.0
PLP[5]	37.6	4.9
PLP[3]	26.7	3.1
MFCC[2]	20.8	5.8
MFCC[3]	17.8	2.7
Shimmer	16.8	4.6

Table 9: Top 10 more common features on explanations.

feature	percentage of subjects	contribution (weight)
MFCC[10]	5.9	7.6
ΔMFCC[1]	5.0	6.8
MFCC[2]	20.8	5.8
PLP[0]	77.2	5.4
$\Delta\Delta$MFCC[0]	1.0	5.3
ΔMFCC[7]	2.0	5.1
MFCC[12]	38.6	5.0
PLP[5]	37.6	4.9
MFCC[0]	65.3	4.8
MFCC[1]	44.6	4.7

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

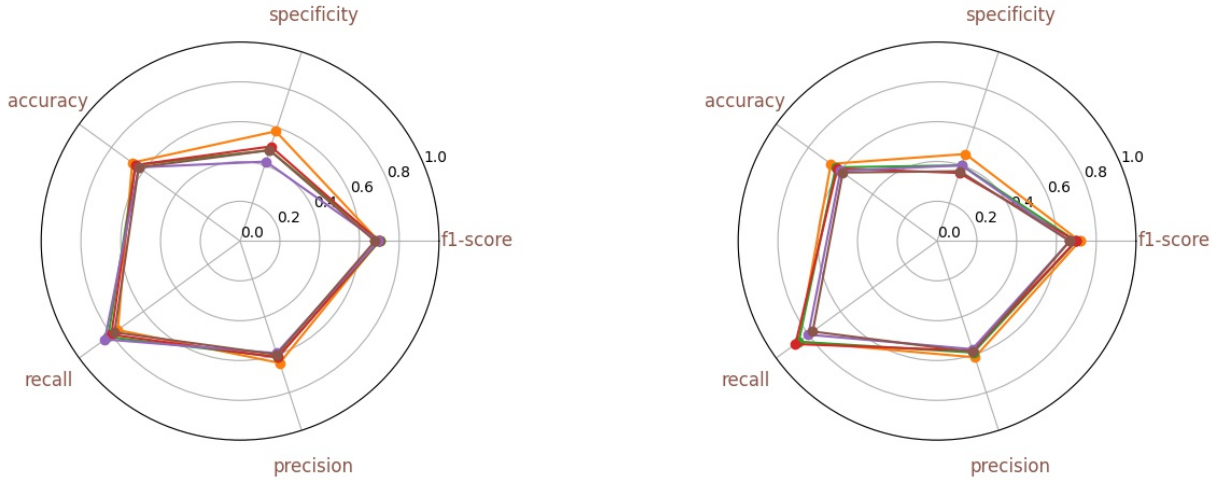

Fig. 4: Independent experiment result using architecture 1 (left) and architecture 2 (right). Correspondence to table 7 and 8: row 1 is orange, row 2 is green, row 3 is red, row 4 is purple, row 5 is brown.

Prediction Probabilities

Healthy	Parkinson
30	70

Feature weight

Feature	PD probability contribution (%)	Subject's average value	Healthy values	Description
pcm_fftMag_mfcc[1]	10.3	0.900		Features that approximate to our perception of the audio quality.
PlpCC[0]	6.9	-1.111		Features that approximate to our perception of the audio quality.
pcm_fftMag_mfcc[0]	-6.1	-3.920		Features that approximate to our perception of the audio quality.
shimmerLocaldB_sma3nz	-6.0	1.378	< 40	The ratio between periodic (associated with normal speed production) and non-periodic (associated with noise) speech components.
PlpCC[2]	5.6	-0.523		Features that approximate to our perception of the audio quality.

Fig. 5: Example explanation report generated by LIME.

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 9 and 10).
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 9 and 10). Comparing the MFCCs and PLPs by percentage of subjects, there are no significant differences between these features. On the other hand, 8 of the 10 features with highest contribution are MFCC parameters. In addition, shimmer is also on the top features ordered by number of subjects for which they are the most relevant. Finally, jitter and F0 produce significant contributions to few test subjects (11.9% jitter and 1% for F0). These features' contributions are inferior to the ones shown on the table (2.9% for F0, and 2% for jitter). HNR was never one of the top five features for any subject.
The global contribution (weight) for each feature can be observed in figure 6. The contribution of two features with lowest weight is significantly smaller than the remaining. In addition, there is a significant difference between the weight of the three features with highest contribution and the others, which can be defined as a threshold to separate the features into two groups (*relevant* and *irrelevant*).
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 6^{th} and the 7^{th} top features (sorted by number of subjects), 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 six features, meaning that this should be the group of features that the medical professional should focus on.

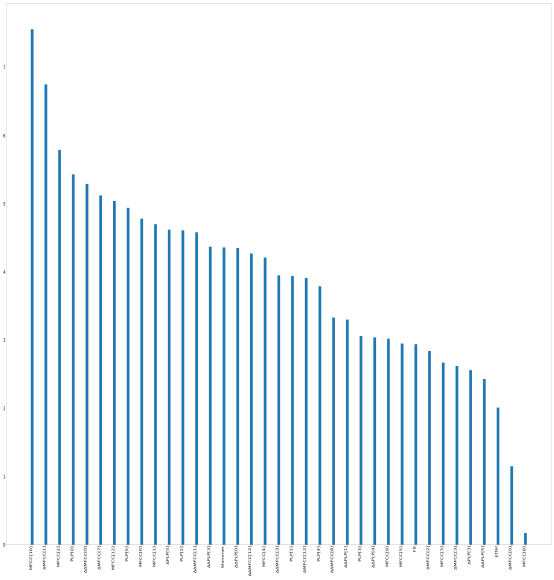

Fig. 6: Global contribution (weight) by feature.

6 Conclusions

This work addressed two issues of the PD diagnosis task: universality and explainability. First, lack of training data for PD diagnostic creates a necessity for pre-trained models that can be re-trained with a small dataset of speech from a new language and be able to diagnose patients speaking the new language. Secondly, lack of ability to understand black-box models' diagnosis is a barrier to real-world usage of such models, which can be solved using explainability models.

6.1 Conclusions

In the first part of this work, we evaluated the performance of a new language-independent model for the PD diagnosis task. Baseline results (training and testing the model with speech from subjects speaking the same language) achieved a maximum accuracy of 90% with the two MLP architectures tested. These architectures differ in the number of layers and number of nodes per layer. An intermediate step was taken between the baseline and a language-independent model, in which models were trained with one dataset and with 90% of another dataset (with different language speakers) and tested with the remaining 10% of the second dataset. Both architectures yielded a maximum accuracy of 90% in the intermediate setting, without loosing performance compared to the baseline. This demonstrated the capacity of these models to be applied to a different language with a smaller amount of training data when pre-trained with a different language. This characteristic can be useful as the size of the available datasets limits the quality of the training. Although results of the present work were very promising, the percentage of the new language dataset used for training (90%) is still high. Reducing the amount of data to re-train a model is worth investigating in the future. When training a language-independent model (trained with two datasets and tested with a different one, from a new language), accuracy dropped to two thirds for both architectures. Results were inferior to the state-of-the-art regarding the accuracy metric, as a similar work achieved an accuracy of 77% with a language-independent model [42]. Our model with highest accuracy yielded a maximum value of 76% on the recall metric, significantly higher than the 53% achieved by the work of Orozco-Arroyave, *et al.* (2016). Therefore, the present work produces less false negatives (PD subjects classified as HC), thus being a more robust tool to support the medical activity. In the second part of the work, the LIME model was used to generate an explanation for each diagnostic. This step allowed to explain the classification results in a way understandable by medical professionals, thus providing trust in the model. This explanation can foster the adoption of computational diagnostic models to be used in clinical scenarios, as these models often produce more accurate diagnostics than medical professionals. The LIME explanation report indicated the probability of each subject belonging to on of the classes (PD and HC), as well as the top five acoustic features which contribute the most to the model's classification. For each feature, the contribution weight and the subject's average value were also included, together with the range of values of a healthy patient and a small description of the feature. This report largely extends the information produced by the classification model, which only indicates the final diagnos-

tic, thus providing the medical professional with information that allows to make an informed diagnostic. Finally, a global analysis was conducted to evaluate the average contribution of each acoustic feature extracted and the percentage of test subjects for which each feature was one of the top five with the highest contribution. Combining both results, we concluded that MFCC and PLP features represent better information for the PD diagnostic task than F0, jitter, shimmer, and HNR. Note that both MFCC and PLP are abstract mathematical representations of sound, and are therefore difficult to explain to a medical professional. Additionally, to the best of our knowledge, there is no known range of values for both MFCC and PLP parameters that defines a healthy patient, which prevents our model to generate a complete report on these features. This work should be extended in the future with simpler, easy to understand features in order for the model to be used in a clinical scenario.

6.2 Future work

There are several 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. Using simpler features, such as Logarithmic Filter Banks [34] (instead of MFCC), would increase the clarity, and therefore the trustworthiness/reliability of the model's diagnostic. 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 range between 105 and 160 *Hz* and between 175 to 245 *Hz* for females). Thus, adding the gender as a feature for the model could help improve the model's performance.
Both the classification and the explanation pipeline's steps can be further improved. First, the similarity between the average contribution (weight) of all features on the explanation model suggests some correlation between features. This hypothesis can be further studied, using a model to evaluate the interactions between features, such as factorization machines [54]. Detecting redundant features could help reduce the model's complexity, thus reducing resource requirements. Also, the results achieved on the semi language-independent experiments showed performance was not affected 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 possible performance loss. Finally, both for the classification and the explanation steps, different models can be used to make a comparative analysis. This would allow to both assess the classification ability of multiple models, and 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 of 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 or images) should also be compared, as to understand which are better accepted by medical professionals. For example, the explanations generated by the model developed during this work could be compared with the 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.

References

1. Rishabh Agarwal, Nicholas Frosst, Xuezhou Zhang, Rich Caruana, and Geoffrey Hinton. Neural additive models: Interpretable machine learning with neural nets. 2020.
2. Razvan Andonie. Extreme data mining: Inference from small datasets. *International Journal of Computers, Communications & Contro*, 3:280–291, 2010.
3. Maral Asiaee, Amir Vahedian-azimi, Seyed Atashi, and Abdalsamad Keramatfar. Voice quality evaluation in patients with covid-19: An acoustic analysis. *Journal of Voice*, 2020.
4. S. Bach, Alexander Binder, Grégoire Montavon, K. Müller, and W. Samek. Analyzing classifiers: Fisher vectors and deep neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2912–2920, 2016.
5. Z. Bakar, N. Tahir, and I. Yassin. Classification of parkinson's disease based on multilayer perceptrons neural network. In *2010 6th International Colloquium on Signal Processing its Applications*, pages 1–4, 2010.
6. Y. Bang, K. Min, Y. Sohn, and S. Cho. Acoustic characteristics of vowel sounds in patients with parkinson disease. *NeuroRehabilitation*, 32:649–654, 2013.
7. B. Bot, C. Suver, E. Neto, Michael Kellen, Arno Klein, Christopher Bare, Megan Doerr, Abhishek Pratap, J. Wilbanks, E. Dorsey, S. Friend, and A. Trister. The mpower study, parkinson disease mobile data collected using researchkit. *Scientific Data*, 3, 2016.
8. Diogo Braga, Ana Madureira, Luis Coelho, and Reuel Ajith. Automatic detection of parkinson's disease based on acoustic analysis of speech. *Engineering Applications of Artificial Intelligence*, 77:148–158, 2019.
9. Irene Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in health care. *The Hastings Center Report*, 2020.
10. Bongjae Choi, Geumsook Shim, Bumseok Jeong, and Sungho Jo. Data-driven analysis using multiple self-report questionnaires to identify college students at high risk of depressive disorder. *Scientific Reports*, 10, 2020.
11. Joon Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. In *Interspeech*, page 2616–2620, 2017.

12. Joon Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *Interspeech*, page 2616–2620, 2018.
13. Mayo Clinic. Parkinson's disease. https://www.mayoclinic.org/diseases-conditions/parkinsons-disease/symptoms-causes/syc-20376055, 2020. Online; accessed 23 December 2020.
14. Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey, 2020.
15. Vladimir Despotovic, Tomas Skovranek, and Christoph Schommer. Speech based estimation of parkinson's disease using gaussian processes and automatic relevance determination. *Neurocomputing*, 401:173–181, 2020.
16. E. Dorsey, Todd Sherer, Michael Okun, and Bas Bloem. The emerging evidence of the parkinson pandemic. *Journal of Parkinson's Disease*, 8:S3–S8, 12 2018.
17. B. Erdogdu Sakar, M. Isenkul, C.O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, and O. Kursun. Collection and analysis of a parkinson speech dataset with multiple types of sound recordings, 2013.
18. Florian Eyben, Martin Wöllmer, and Björn Schuller. opensmile – the munich versatile and fast open-source audio feature extractor. pages 1459–1462, 2010.
19. Joana Fernandes, Felipe Teixeira, Vitor Guedes, Arnaldo Junior, and João Paulo Teixeira. Harmonic to noise ratio measurement - selection of window and length. *Procedia Computer Science*, 138:280–285, 2018. CENTERIS 2018 - International Conference on ENTERprise Information Systems / ProjMAN 2018 - International Conference on Project MANagement / HCist 2018 - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2018.
20. Wojciech Froelich, Krzysztof Wróbel, and Piotr Porwik. Diagnosis of parkinson's disease using speech samples and threshold-based classification. *Journal of Medical Imaging and Health Informatics*, 5:1358–1363, 2015.
21. Alexander Goberman and Lawrence Elmer. Acoustic analysis of clear versus conversational speech in individuals with parkinson disease. *Journal of Communication Disorders*, 38:215–230, 2005.
22. J. Godino-Llorente, S. Shattuck-Hufnage, J. Choi, Laureano Moro-Velázquez, and J. Gómez-García. Towards the identification of idiopathic parkinson's disease from the speech. new articulatory kinetic biomarkers. 2017.
23. Brian Harel, Michael Cannizzaro, Henrí Cohen, Nicole Reilly, and Peter Snyder. Acoustic characteristics of parkinsonian speech: a potential biomarker of early disease progression and treatment. *Journal of Neurolinguistics*, 17:439–453, 2004.
24. M. Hely, W. Reid, M. Adena, G. Halliday, and J. Morris. The sydney multicenter study of parkinson's disease: the inevitability of dementia at 20 years. *Movement Disorders Journal*, 23:837–844, 2008.
25. H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. Rasta-plp speech analysis technique. In *ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 121–124, 1992.
26. Margaret M. Hoehn and Melvin D. Yahr. Parkinsonism. *Neurology*, 17(5):427–427, 1967.
27. Andreas Holzinger, Chris Biemann, Constantinos Pattichis, and Douglas Kell. What do we need to build explainable ai systems for the medical domain? 2017.
28. Hagen Jaeger, Dhaval Trivedi, and Michael Stadtschnitzer. 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], 2019.
29. T. Kent, H. Vorperian, J. Kent, and J. Duffy. Voice dysfunction in dysarthria: application of the multi-dimensional voice program. *Journal of Communication Disorders*, 36:281–306, 2003.
30. M. Little, P. McSharry, D. Costello, and I. Moroz. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMedical Engineering OnLine*, 2007.
31. Alex London. Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *The Hastings Center Report*, 49:15–21, 2019.
32. Karmele Lopez-de Ipiña, Jordi Solé-Casals, Harkaitz Eguiraun Martinez, Jesús Alonso, Carlos Travieso, M. Ecay, Aitzol Ezeiza, Nora Barroso, Pablo Martinez-Lage, and Blanca Beitia. Feature selection for spontaneous speech analysis to aid in alzheimer's disease diagnosis: A fractal dimension approach. *Computer Speech & Language*, 30, 2014.
33. Pavan Magesh, Richard Myloth, and Tom Rijo. An explainable machine learning model for early detection of parkinson's disease using lime on datscan imagery. *Computers in Biology and Medicine*, 126, 2020.
34. W.F. McGee and G. Zhang. Logarithmic filter banks. In *IEEE International Symposium on Circuits and Systems*, pages 661–664 vol.1, 1990.
35. A. Moreno, D. Poch-Olivé, A. Bonafonte, E. Lleida, J. Llisterri, J. Mariño, and C. Nadeu. Albayzin speech database: design of the phonetic corpus. In *Eurospeech*, 1993.
36. Laureano Moro-Velazquez, Jorge Gomez-Garcia, Juan Godino-Llorente, and Najim Dehak. 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, 2019.
37. Laureano Moro-Velázquez, Jorge Gomez-Garcia, Juan Godino-Llorente, Francisco Grandas-Perez, Stefanie Shattuck-Hufnagel, Virginia Yagüe-Jimenez, and Najim Dehak. Phonetic relevance and phonemic grouping of speech in the automatic detection of parkinson's disease. *Scientific Reports*, 9, 2019.
38. Laureano Moro-Velázquez, Jesus Villalba, and Nasim Dehak. Using x-vectors to automatically detect parkinson's disease from speech. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1155–1159, 2020.
39. Lizbeth Naranjo, Carlos Pérez, Yolanda Campos-Roca, and Jacinto Martín. Addressing voice recording replications for parkinson's disease detection. *Expert Systems With Applications*, 46, 2016.
40. National Institute of Aging. Parkinson's disease. https://www.nia.nih.gov/health/parkinsons-disease, 2017. Online; accessed 23 December 2020.
41. J. Orozco-Arroyave, J. Arias-Londoño, J. Vargas-Bonilla, M. Gonzalez-Rátiva, and E. Nöth. New spanish speech corpus database for the analysis of people suffering from parkinson's disease. In *LREC*, 2014.
42. Juan Orozco-Arroyave, Florian Honig, Julian Arias-Londoño, J. Francisco Vargas-Bonilla, Khaled Daqrouq, Sabine Skodda, Jan Rusz, and Elmar Noth. Automatic detection of parkinson's disease in running speech spoken in three different languages. *Journal of the Acoustical Society of America*, 139:481–500, 2016.
43. Fernando Pagan. Improving outcomes through early diagnosis of parkinson's disease. *American Journal of Managed Care*, 18, 2012.

44. Arrigo Palumbo, Barbara Calabrese, P. Vizza, Nicola Lombardo, Aldo Garozzo, Mario Cannataro, Francesco Amato, and Pierangelo Veltri. *A Novel Portable Device for Laryngeal Pathologies Analysis and Classification*, volume 55, pages 335–352. 01 2010.
45. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
46. K. Perez, L. Ramig, M. Smith, and C. Dromey. The parkinson larynx: tremor and videostroboscopic findings. *Journal of voice : official journal of the Voice Foundation*, 1996.
47. Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *CoRR*, 2018.
48. André Pfob, Chris Sidey-Gibbons, Han-Byoel Lee, Marios Konstantinos Tasoulis, Vivian Koelbel, Michael Golatta, Gaiane Rauch, Benjamin Smith, Vicente Valero, Wonshik Han, Fiona MacNeill, Geraldine Rauch, and Henry Kuerer. 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, 2020.
49. Serge Pinto, Rita Cardoso, Jasmin Sadat, Isabel Guimarães, Céline Mercier, Helena Santos, Cyril Atkinson-Clement, Joana Carvalho, Pauline Welby, Pedro Oliveira, Mariapaola D'Imperio, Sonia Frota, Alban Letanneux, Marina Vigário, Marisa Cruz, Isabel Martins, François Viallet, and Joaquim Ferreira. 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, 2016.
50. Anna Pompilli, Alberto Abad, Paolo Romano, Isabel Pavão Martins, Rita Cardoso, Helena Santos, Joana Carvalho, Isabel Guimarães, and Joaquim Ferreira. 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, 2017.
51. J. Proença, F. Perdigão, A. Veira, S. Candeias, J. Lemos, and C. Januário. Characterizing parkinson's disease speech by acoustic and phonetic features. *Computational Processing of the Portuguese Language*, pages 24–35, 2014.
52. Narinder Punn and Sonali Agarwal. Automated diagnosis of covid-19 with limited posteroanterior chest x-ray images using fine-tuned deep neural networks. *Applied Intelligence*, pages 1–14, 2020.
53. Mahendra Rai, Alka Yadav, Avinsh P. Ingle, Anatoly Reshetilov, María José Blanco-Prieto, and Chistiane Feitosa. *Neurodegenerative Diseases: The Real Problem and Nanobiotechnological Solutions*, pages 1–17. Springer International Publishing, 2019.
54. Steffen Rendle. Factorization machines. In *2010 IEEE International Conference on Data Mining*, pages 995–1000, 2010.
55. Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "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*, page 1135–1144, 2016.
56. J. Rusz, R. Cmejla, T. Tykalová, H. Ruzickova, J. Klempír, V. Majerova, J. Picmausova, J. Roth, and E. Růika. Imprecise vowel articulation as a potential early marker of parkinson's disease: effect of speaking task. *The Journal of the Acoustical Society of America*, pages 2171–2181, 2013.
57. Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. *CoRR*, 2017.
58. John Tracy, Yasin Özkancab, David Atkins, and Reza Ghomi. Investigating voice as a biomarker: Deep phenotyping methods for earlydetection of parkinson's disease. *Journal of Biomedical Informatics*, 104, 2020.
59. A. Tsanas, M. Little, P. McSharry, and L. Ramig. Accurate telemonitoring of parkinson's disease progression by non-invasive speech tests. *IEEE Transactions on Biomedical Engineering*, 2009.
60. Ole-Bjørn Tysnes and Anette Storstein. Epidemiology of parkinson's disease. *Journal of neural transmission*, 124:901–905, 2017.
61. Alfredo Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, 2019.
62. Shaohua Wan, Yan Liang, Yin Zhang, and Mohsen Guizani. Deep multi-layer perceptron classifier for behavior analysis to estimate parkinson's disease severity using smartphones. *IEEE Access*, 2018.
63. C. Williams-Gray, S. Mason, J. Evans, T. Foltynie, C. Brayne, T. Robbins, and R. Barker. 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, 2013.
64. Orhan Yaman, Fatih Ertam, and Turker Tuncer. Automated parkinson's disease recognition based on statistical pooling method using acoustic features. *Medical Hypotheses*, 135, 2020.

Glossary

AD Alzheimer's Disease. 1, 4
AI Artificial Intelligence. 1, 2, 4
ARD Automatic Relevance Determination. 3

BC Boosting Classifier. 3

DDK Diadochokinesia. 2, 3, 6
DL Deep Learning. 4
DMLP Deep Multi-Layer Perceptrons. 4
DNN Deep Neural Networks. 4
DT Decision Trees. 4

F0 Fundamental Frequency. 2–4, 7, 14, 16

GMM-UBM Gaussian Mixture Model - Universal Background Model. 3, 4
GNR Glottal-to-Noise Ratio. 2, 3
GPC Gaussian Process Classification. 2, 3

HC Healthy Controls. 2–4, 6–9, 11, 12, 15
HNR Harmonics-to-Noise Ratio. 2, 3, 7, 14

KNN K-Nearest Neighbours. 3, 4

LIME Local Interpretable Model-agnostic Explanation. 4, 5
LR Logistic Regression. 4
LRP Layer-wise Relevance Propagation. 4

MFCC Mel-frequency cepstral coefficients. 3, 14, 16
ML Machine Learning. 1–5
MLP Multi-Layer Perceptrons. 3, 4, 7–9, 15

NAM Neural Additive Models. 5
NHR Noise-to-Harmonics Ratio. 2, 3
NN Neural Networks. 3, 5

PD Parkinson's Disease. 1–9, 11, 12, 15, 16
PLDA Probabilistic Linear Discriminant Analysis. 3
PLP Perceptual Linear Predictive. 7, 14, 16

Rasta-PLP Rasta-Perceptual Linear Predictive. 3
RF Random Forests. 3, 4
RISE Randomized Input Sampling for Explanation. 5

SVM Support Vector Machines. 3

TDU Text-dependent Utterances. 2, 3

VTI Voice Turbulence Index. 2, 3

XAI Explainable Artificial Intelligence. 2, 4