NOVA IMS Information Management School

MDSAA

Mestrado em

Data Science and Advanced Analytics

Parkinson's Disease and Atypical Parkinsonism in the context of rehabilitation:

a machine learning approach

Susana Cristina Norberto Pires

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Parkinson's Disease and Atypical Parkinsonism in the context of rehabilitation:

a machine learning approach

by

Susana Cristina Norberto Pires

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

Supervised by

Professor Leonardo Vanneschi, PhD, NOVA Information Management School

July, 2025

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 15th July 2025

Susana Cristina Norberto Pires

DEDICATION

Para a minha avó, Birrinha ♡

ACKNOWLEDGEMENTS

To my supervisor, Professor Leonardo Vanneschi, I am thankful for all the guidance, support and trust. Even in the most challenging times, there was always word of encouragement.

To the person that had to deal with every moment of this journey, my best friend, my boyfriend, my husband, Filipe Alves. Thank you for all the patience during the challenging moments, for your reassurance when I needed it and for never doubting me.

To my beloved family, I am grateful for all your support, love and understanding throughout this quest.

To my longtime friends, Catarina de Carvalho and Carolina Silva, thank you for your friendship and trust, and patience in dealing with my rambling conversations.

To my friends of this journey, Sofia Pereira and Mariana Cabral, I am thankful for our growing friendship and the support we have given each other along this path.

ABSTRACT

The increasing use of artificial intelligence in healthcare offers significant opportunities to improve patient outcomes and optimize clinical workflows. From the diversity of applications, machine learning continues to demonstrate its potential of use as an auxiliary diagnostic tool. It is used in the context of neurodegenerative diseases research, for tasks such as speech and handwriting analysis, gait impairment studies using sensor data, and to discover new potential pharmacotherapies. Parkinson's disease and atypical parkinsonism are two classes of neurodegenerative pathologies that present significant challenges in differential diagnosis and require distinct management strategies. Distinguishing them early allows for a better prognosis by targeting the specific characteristics of each disease. Despite the advancements in machine learning research applied to Parkinson's, there is a gap in studies that specifically address the binary classification between the two classes, which is the focus of this thesis. This research can serve as a starting point to the development of a complementary tool to assist practitioners with early differential diagnosis, using data from standardised clinical assessments. A comprehensive experimental design was implemented, using six machine learning classifiers. One of the tested implementations accounts for strategies to address class imbalance and small sample size. The objectives were to analyse and identify the best-performing model, and to determine which features impact classification the most. The results, when accounted for F1-score macro, indicate that extreme gradient boosting and random forest achieved the highest scores. Statistical testing revealed no significant difference in performance between these two models. When balanced accuracy was considered, logistic regression, support vector classifier and extreme gradient boosting achieved the highest scores and also presented no statistically significant differences in performance. Although the performance metrics were not as good as expected, the best models achieved scores ranging from 0.55 to 0.62 for both F1-score macro and balanced accuracy. This research is a foundation for the classification task of recognizing each disease. Additionally, model interpretability was explored using Shapley additive explanations. The feature "mds_updrs_pIII" was found to be the most influential in distinguishing between the two conditions, followed by the "Age". These insights can inform future scientific research. Despite limitations in data quality and sample size, this work provides a stable methodology for future studies with richer datasets, more features, and alternative modelling approaches. Ultimately, this research emphasises the value of interdisciplinary collaboration and how both positive and negative results contribute to scientific progress.

KEYWORDS

Machine Learning; Artificial Intelligence; Parkinson's Disease; Atypical Parkinsonian Disorders; Atypical Parkinsonism; Healthcare; Rehabilitation

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ANN	Artificial Neural Network
AP	Atypical Parkinsonism
APD	Atypical Parkinsonian Disorder
BAcc	Balanced Accuracy
BEST	Balance Evaluation Systems Test
CGI	Clinical Global Impression
csv	comma-separated values
cv	Cross Validation
df	DataFrame
DT	Decision Tree
EOPD	Early-onset Parkinson's Disease
ESAS	Edmonton Symptom Assessment System
F1M	Macro-averaged F1 score
F1W	Weighted-average F1 score
GDPR	General Data Protection Regulation
GS	Grid Search
HC	Healthy Control
HY	Hoehn and Yahr Scale
IMU	Inertial Measurement Unit
kNN	k-Nearest Neighbours
LR	Logistic Regression
MCI	Mild Cognitive Impairment
MDS-UPDRS	Movement Disorder Society – Unified Parkinson's Disease Rating Scale
ML	Machine Learning

MLP	Multi-layer Perceptron
MMSE	Mini Mental Examination State
MoCA	Montreal Cognitive Assessment
NaN	Not a Number
NB	Naïve Bayes
N-FoG	New Freezing of Gait (questionnaire)
NN	Neural Network
PD	Parkinson's Disease
PGI	Patient Global Impression
PhT	Pharmacological therapies
PwAP	Patients with Atypical Parkinsonian Disorders
PwPD	Patients with Parkinson's Disease
RF	Random Forest
SE	Schawb & England Scale
SHAP	Shapley Additive Explanations
SkF	Stratified k-Fold
SMOTE	Synthetic Minority Oversampling Technique
std	Standard Deviation
SVC	Support Vector Classifier
SVM	Support Vector Machine
UPDRS	Unified Parkinson's Disease Rating Scale
WHO	World Health Organization
WHODAS	World Health Organization Disability Assessment Schedule
XGB	Extreme Gradient Boosting

1. Introduction

1.1. AI IN HEALTHCARE

The exponential growth of Artificial Intelligence (AI) innovations in recent years has been evident across all sectors, and the healthcare domain is no exception (Hirani et al., 2024). The diversity of solutions covers all levels of complexity. These range from predicting the recurrence of emergency room visits and computer-aided diagnosis software, to entering the area of computer vision and large language models. The impact is undeniable, and the goal is to keep improving the patient outcomes and redistribute the workload of healthcare workers (Khan & Sherani, 2025). However, due to the rigorous legislation and ethical concerns surrounding human health matters, the development of such innovations takes more time than in other sectors (Mohsin Khan et al., 2025). Bearing this in mind, the aspiration of this thesis is to take a small first step towards developing an auxiliary diagnostic tool.

1.2. PARKINSON'S DISEASE AND ATYPICAL PARKINSONISM

According to the Global Burden of Disease in 2021, neurological diseases had a global prevalence of more than 2 870 million cases, including more than 11 million cases of Parkinson's Disease (PD) (University of Washington, 2024). In recent years, there has been an increase in prevalence, which goes hand in hand with the increase in life expectancy (J. Zhang et al., 2024). As a result, the economic burden of this disease is quite significant, and increases substantially with the progression of the disease (Chaudhuri et al., 2024).

Atypical Parkinsonism (AP) presents some similarities with PD, but its distinction is important to a proper disease management and prognostic value (Li et al., 2024). As the results on the most recent advancements do not yet demonstrate clinical and academic agreement, both diseases are still diagnosed based on clinical features (Munhoz et al., 2024; Virameteekul et al., 2023). Two of the most relevant ones are the duration of symptoms and its rapid progression, and the poor response to levodopa, that is common therapeutic drug for PD symptom management (Cardoso et al., 2024).

1.3. MACHINE LEARNING IN NEURODEGENERATIVE DISEASES

The application of Machine Learning (ML) models dedicated to developments in the field of neurology and neurodegenerative diseases has been increasing over the past years. And regarding PD, there are several approaches dedicated to different aspects of the disease (Gupta et al., 2023). These include tasks such as analysing speech (Devarajan et al., 2023) and handwriting for early diagnosis (Kamran et al., 2021; Thomas et al., 2017), studying gait impairments (Pogrzeba et al., 2019) with sensor data (Facciorusso et al., 2023), and

accelerating the process of discovering new pharmacological therapies (Dara et al., 2022; Visanji et al., 2021). These studies and similar others reflect an important impact on what is known about PD and other neurodegenerative diseases, with some even offering new tools to assist the medical practitioners on the early identification of PD distinctive clinical features.

1.4. MOTIVATION AND OBJECTIVES

As in any other health-related work, the motivation behind this project is clear: to improve patient outcomes, in this case by using ML techniques. As there is currently no known cure for either PD or AP, the early distinction allows for a more targeted disease management, with the aim of improving the patients' quality of life (Lee & Yankee, 2022).

Regarding the research gap, during the literature review, which focused on studies using ML on data from patients with neurodegenerative disorders, no study was found that focused on the distinction between patients with PD (PwPD) and patients with AP (PwAP) using rehabilitation data (physiotherapy, nursing, neuropsychology, speech therapy and others). Based on this specific data, two research objectives have been formulated. One is to determine a ML model that best classifies whether a patient has PD or APD. The second if to identify which features have the most influence on the previous classification.

One of the expected results is to find a model with robust performance metrics on the binary classification of distinguishing PD from AP. With a high performing model, it is possible to aim for future research and development of an auxiliary tool for clinicians. This would also benefit the patients by improving their quality of life, as diagnosis would be faster, consequently improving the disease management outcomes. Another expected result is to identify the features that most impact the models' classification, which allow for better interpretation of the results and potentially leading to new clinical research focused on those features.

1.5. DOCUMENT STRUCTURE

This thesis is organized into six chapters. The current Chapter 1 provides an introduction to this project and an overview of the thesis. Chapter 2 presents the background concepts to provide proper contextualisation of all the subjects explored. Chapter 3 reviews the scientific literature and studies on similar topics and supports the identification of the research gap. Chapter 4 details the methodological approach to each of the proposed objectives, providing comprehensive information on every step of the project, from initial data access to the final model implementation. Chapter 5 presents the results and scrutinises them to ensure correct interpretation, comparing different models and performance metrics while conducting statistical testing. Chapter 6 concludes the answers to the research objectives based on the data from the previous chapter and explores the limitations of this study while promoting new ideas for future research.

2. Background

2.1. Machine learning concepts

Machine Learning (ML) can be defined as a subfield or branch of Artificial Intelligence (AI), although it is also its own interdisciplinary field, as encompasses knowledge from other domains such as mathematics and statistics (Shalev-Shwartz & Ben-David, 2014). The objective of ML is to develop programs that successfully learn to solve complex tasks (Mitchell, 1997). The developments of ML in the healthcare sector have been growing exponentially, and some even surpass the capacities of highly experienced professionals (Topol, 2019). However, these advancements are decelerating since it takes time to develop these tools according to the most rigorous protocols of the health industry (Krones et al., 2025).

As in any field of research, some common guidelines are needed to achieve reproducible results. In this project, the selected framework to deal with the data is the Cross Industry Standard Process for Data Mining (CRISP-DM), and the structure of this iterative process is set up by six phases (Figure 1).

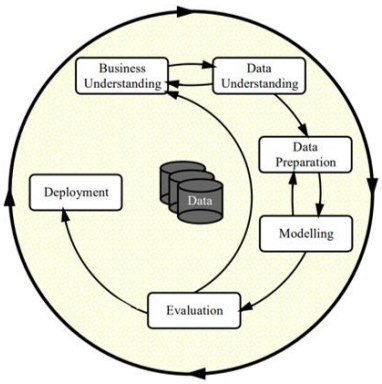

Figure 1 - Phases of the CRISP-DM Process Model for Data Mining (Wirth & Hipp, 2000)

The main paper on this process states that each phase has its own tasks and outputs (Wirth & Hipp, 2000). While other studies provide recommendations on how to break down each phase into a list of tasks and expected outputs (Schröer et al., 2021).

For this project, six ML models were experimented, based on similar procedures found in the literature. Five out of the six models are tested with the scikit-learn implementation, and only the XGBoost has its own library. Logistic Regression (LR) was selected as a baseline approach. It is a statistical model that is implemented as a supervised ML algorithm for classification tasks (Matloff, 2017), using the scikit-learn implementation (scikit-learn: LR, 2025). The LR has the advantages of its computational efficiency and interpretability, hence defining it as the baseline for comparison of the other models.

The following model is Naïve Bayes (NB) (scikit-learn: NB, 2025). This model applies the Bayes' theorem, but with the *'naïve'* assumption of conditional independence of every pair of features (H. Zhang, 2004). Although its apparent simplicity, it has achieved good performance in real-life tasks, and its use is reported in similar literature. Another model is the Random Forest Classifier (RF) (scikit-learn: RF, 2025), which is a combination of tree predictors fitted on subsamples of the dataset and they are more robust to overfitting (Breiman, 2001).

Other model is the Multi-layer Perceptron (MLP) (scikit-learn: MLP, 2025). It is a feedforward artificial neural network (ANN) trained using backpropagation, suitable for complex classification tasks (Kruse et al., 2022). While Support Vector Classification (SVC) (scikit-learn: SCV, 2025) is the fifth selected model, that tends to find an optimal hyperplane to separate the classes with a maximum margin, although it is appropriate for high-dimensional spaces, it still reports good performance on smaller sets (Chang & Lin, 2011).

The final experimented model is the eXtreme Gradient Boosting (XGB) under its Python library implementation (eXtreme Gradient Boosting, 2025). The XGB is used in a diversity of problems since it is an optimized version of a scalable tree boosting system (Chen & Guestrin, 2016).

The metrics selected to evaluate the models are balanced accuracy (BAcc) (scikit-learn: bacc, 2025) and macro F1-score (F1M) (scikit-learn: f1, 2025), since they are suitable to deal with imbalanced datasets (Diallo et al., 2025). The BAcc provides a fair estimation of the model performance, as it mitigates the biases caused by class imbalance, since it considers specificity and sensitivity (Brodersen et al., 2010). The F1M balances precision and recall measures independently for each label, and the macro component allocates an equal weight to each class (regardless of the number of cases in each class) (Hinojosa Lee et al., 2024). For academic purposes, the weighted F1-score was also observed in some cases, because it was reported on some of the articles reviewed. This calculates the metrics for each class and averages them, accounting for the imbalance, but as a result it gives less importance to the minority class.

Regarding statistical testing to determine the best models, Friedman test is used to identify any differences between the six models tested. It is a non-parametric test applied to compare multiple models, across different cross-validation (cv) folds (Liu & Xu, 2022). If the Friedman test rejects the null hypothesis (H_0: There is no significant difference in performance between the models), then a *post-hoc* analysis needs to be conducted (Benavoli et al., 2016). This analysis can be done using different tests such as the Wilcoxon tests with Bonferroni or Holm corrections, or the Nemenyi test (Pereira et al., 2015). The selection of one of these tests will depend on the researcher focus regarding error interpretation.

As for the evaluation of feature importance on the best model, Shapley Additive Explanations (SHAP) (*SHAP*, 2025) is a method for interpreting ML models by quantifying the contribution of each feature to a specific prediction. It helps to define an interpretable approximation of the complex models, and works as a unified measure of feature importance (García & Herrera, 2008; Lundberg & Lee, 2017).

2.2. HEALTH, DISEASE AND REHABILITATION CONCEPTS

PD is a progressive neurodegenerative disorder, characterized by four main features: tremor at rest, rigidity, akinesia and postural instability (Jankovic, 2008). The clinical diagnosis of PD is centred on these defined motor syndromes since its original description. However, diagnostic certainty is not possible during life, as the gold-standard for PD diagnosis is through analysing *post-mortem* brain tissue (Postuma et al., 2015). Over the years, some non-motor features have been described in the literature (Smith et al., 2021) and it has been helpful to improve the neurologists clinical expertise in determining PD diagnoses. Typically, PD is diagnosed on older patients, but when patients are under 50 years old, it is diagnosed as an early-onset Parkinson's disease (EOPD) as it has different characteristics and prognosis (Mehanna et al., 2022).

Atypical Parkinsonian disorders (APD) refer to a heterogeneous group of progressive diseases. These include conditions such as: corticobasal degeneration (CD), dementia with Lewy bodies (DLB), multiple system atrophy (MSA) and progressive supranuclear palsy (PSP) (Höglinger et al., 2017). In the context of this project, these will be referred to as AP (atypical parkinsonism). The differential diagnosis between AP and PD is a challenging task, as they share common symptoms and characteristics, such as bradykinesia (Bologna et al., 2023). Therefore, the majority of diagnoses change over time as some symptoms become more noticeable than others (Deutschländer et al., 2018). Some of the possible differences are the duration of the disease and the response to levodopa (Grażyńska et al., 2020).

As there is no cure for any of these pathologies, the rehabilitation approach and disease management have a greater impact on patients' quality of life (Lee & Yankee, 2022). Physical examinations and assessments of motor and non-motor functions are carried out by professionals from various domains, including physiotherapy, occupational therapy, nursing and neurology. Throughout the literature, certain assessment names appeared more frequently as features used to train ML models. Some of these are: MDS-UPDRS the acronym for movement disorder society – unified Parkinson's disease rating scale (Goetz et al., 2008); sit-to-stand (Derby Hospitals NHS Foundation Trust, 2003); MMSE: mini-mental state examination (Tombaugh & McIntyre, 1992); HY: Hoehn and Yahr scale (Martinez-Martin et al., 2018); MoCA: Montreal cognitive assessment (Nasreddine et al., 2005); and timed up and go (Keus et al., 2014).

3. Literature Review

3.1. Literature search and selection methods

Considering that the available dataset mainly contains data relating to rehabilitative physiotherapy assessments of patients with Parkinson's disease (PwPD) and patients with atypical parkinsonism (PwAP), the literature search began with a very specific query.

In January of 2025, Scopus database was checked using the following query:

(TITLE-ABS-KEY ("parkinson" OR "parkinson's" OR parkinsonian OR parkinsonism) OR TITLE-ABS-KEY (neurodegenerative) AND TITLE-ABS-KEY ("artificial intelligence" OR "machine learning" OR "deep learning") AND TITLE-ABS-KEY (physiotherapy OR "physical therapy" OR rehabilitation OR rehab))

This resulted in finding 278 documents. The results were selected by reading the title and the available abstract. If the content seemed relevant to this thesis, the reference was exported to Zotero (reference manager), resulting in a library of 57 items (articles, reviews and conference papers). From there, each article was skimmed through, and relevant information was noted on a broad literature review table for ease of organisation. In the process of reading the full articles, 45 new relevant references were found, mainly from citations in reviews. In terms of exclusion criteria, documents written in a language other than English or Portuguese were not included, nor were articles that cannot be accessed through the Sophos Connect NOVA IMS VPN.

3.2. Diversity of machine learning applications for Parkinson's disease

The variety of ML models dedicated to different aspects of PD includes tasks such as speech and handwriting analysis for early diagnosis, measuring gait impairment using motion and sensor data, and speeding up the process of discovering new therapeutic drugs (Gupta et al., 2023).

Devarajan et al. (2023) point out the importance of how an ML approach can work as a data-driven assistant in the medical field, not aiming to replace the clinicians, but to reduce the diagnostic errors. They used voice data and tested artificial neural networks (ANN), support vector machine (SVM), decision tree (DT) and random forest (RF). Since their study was designed to obtain a model with diagnostic performance, they developed some hybrid ensembles, the best of which was the RF-ANN, with a sensitivity of 98.15%.

Although it is not included as a diagnostic criterion for PD, the decline of handwriting abilities is often observed (Thomas et al., 2017). The work of Kamran et al. (2021) presents an approach using handwriting samples of PwPD for early diagnosis. They use deep transfer learning methods (AlexNet, GoogleNet, VGG-16, VGG-19, ResNet50 and ResNet101) combined with

data augmentation techniques to achieve the best results. However, it is suggested to improve the models by adding clinical features to the data.

On the molecular design and therapeutics side, ML accelerates the discovery process (Dara et al., 2022). The work of Visanji et al. (2021) used ML to predict which of the existing drugs would have a specific effect on a known disease pathway of PD.

In terms of gait assessment parameters, these range from qualitative and quantitative questionnaires administered by professionals with some degree of subjectivity (Pogrzeba et al., 2019), to calibrated quantitative data recorded by sensors. These wearable medical devices can generate immense amounts of new data, which are extremely useful for improving ML algorithms towards personalised rehabilitation solutions and earlier diagnosis (Facciorusso et al., 2023).

Russo et al. (2024) used kinetic sensor data and supervised ML to identify gait features in mild cognitive impairment (MCI) in PD. This study consisted of using data without a healthy control (HC) group, as they concluded it might make the differences between the two groups (PwPD with MCI and PwPD without MCI) clearer. They experimented with seven ML classifiers: DT, RF, gradient boosted tree, SVM, k-Nearest Neighbours (kNN), linear discriminant analysis (LDA) and Naïve-Bayes (NB), and the tree-based algorithms achieved the higher metrics.

3.3. RELATED RESEARCH

Bearing in mind the objectives of this thesis: to find an ML model that is able to classify subjects between suffering from PD or APD and to identify which features have the most impact on this distinction. This subchapter will focus on articles with similar projects, or at least with some common approaches and metrics for comparison, as summarized on Table 1.

Table 1 - Literature Review

Author	Year	Title	Objectives and Conclusions	Data and size	Models	Best model and metrics
D. Jacob, et al.	2022	Assessing Early-stage Parkinson's Disease Using BioVRSera	To predict using ML, based on experimental measurements, whether a subject belonged to healthy or Parkinson's group and extract most important features. Prototype of quantitative evaluation of early-stage Parkinson's.	Electromyography (EMG); Heart Rate (HR); Centre of Pressure (CoP) features. 11 early-stage PwP (patients with Parkinson's) + 46 HC (healthy control)	Tree Medium; Logistic Regression (LogReg); Gaussian Naïve Bayes; Weighted kNN; Subspace Discriminant; Medium Neural Network	Weighted kNN and Medium NN – both with 94.6% accuracy
J. Templeton, et al.	2022	Classification of Parkinson's disease and its stages using machine learning	Classification of disease stage (early vs advanced). To help increase efficacy for diagnostic and rehabilitative purposes.	Neurocognitive functional tests, functional movement assessments and standardized health questionnaires. 50 PwP + 25 HC	Classification and Regression Tree (CART)	Mean value of both classes for each metric (accuracy, precision, recall), evaluated per type of assessment (table 3).
V. Tsakanikas, et al.	2023	Evaluating Gait Impairment in Parkinson's Disease from Instrumental Insole and IMU Sensor Data	Identify the capacity of different sensor features to train ML models to predict gait impairment (binary).	Insole and IMU sensor (kinematic) data on 5 specialized tests. 19 PwP	Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting, AdaBoost	SVM outperformed the other models in all three analyses, considering AUC. Other metrics: accuracy, F1-Score, precision, recall
Y. Han, et al.	2023	Automatic Assessments of Parkinsonian Gait with Wearable Sensors for Human Assistive Systems	To objectively assess gait task in UPDRS with wearable sensors data, using ML.	IMU data: 12 gait features, 3 spatial-temporal features, 9 kinematics features. 25 PwP + 28 HC	Proposed nonlinear model, SVM, Naïve Bayes (NB), Multiple Linear Regression.	Proposed nonlinear model shows higher accuracy (84.9%)
A. Khamparia, et al.	2024	Cognitive driven gait freezing phase detection and classification for neuro-rehabilitated patients using machine learning algorithms	ML classification between gait disorder and typical walking. Experiment how different dimensionality reduction techniques impact the models' performance.	Daphnet Freezing of Gait Dataset (data from acceleration sensors on hip and leg). 237 instances + 9 attributes	LogReg, SVM, RF, DT, kNN, NB, Perceptron	kNN and RF were the ones achieving higher metrics (accuracy, precision, recall, F1-Score).

Based on the study of Jacob et al. (2022), a group of 46 healthy individuals and 11 early-stage PwP was used. The data collected from this group gathers Centre of Pressure (or sway) characteristics and the neurophysiological attributes of electromyography and heart rate while they experience the BioVRSea - a virtual reality experiment set up. Their goal was to efficiently use an ML algorithm to predict whether a subject undergoing this experiment is healthy or suffers from PD. Although they refer the small dataset of 11 PwP as a limitation, they sought to identify the features that distinguish the two groups, in the hope of providing new insights into targeted rehabilitation and early diagnosis. In the end, Jacob et al. concluded that the participants' physiological response was different between the two groups and the best performing ML models were weighted-kNN and medium NN, both with an accuracy of 94.6% in distinguishing between the two groups of individuals.

Templeton et al. (2022) present a paper on the stage classification of PD, with the aim of using supervised ML classification algorithms to experiment with new features collected from digital neurocognitive assessments. Due to small number of participants (50 PwP + 25 HC), data from self-reported metrics and clinically relevant functional movement analysis were also added to the process in the hope of identifying objective features for both disease and stage classification. The model used for this study was the Classification and Regression Tree (CART), which achieved accuracies between 76.32% and 100% for the classification of PwP and HC; and accuracies between 68.42% and 89.47% for the staging classification (adapted from 5 levels of the Hoehn and Yahr scale to two stages: early and advanced). Finally, they point out the importance of experimenting with additional ML (such as kNN, DT, LogReg, RF), since PD is a disease for which personalized treatment makes sense.

In the paper of Tsakanikas et al. (2023) the use of ML methods focused on the assessment of gait impairment using objective sensor data (insole and kinematic IMU). First, they explored the correlation between the features to identify the training capability for the later ML model. Then, for the binary classification task, four ML models (SVM, RF, GB, AdaBoost) were tested, each with three sub-analyses corresponding to the different feature sets. This study concluded, based on the AUC values, that SVM outperformed the other models in all three analyses. To complete the objectives, they used Shapley Additive Explanations (SHAP) to evaluate the importance of each feature to the prediction output.

Although the UPDRS is a standard clinical evaluation tool for PwPD, consisting of five levels reflecting the gait task scores, it is a type of assessment that leaves some subjectivity to the experience of the neurologist. Thus, Han et al. (2023) propose an objective and automatic assessment of gait in PwPD using data from wearable sensors. They consider the analysis of 24 features (9 kinematic, 3 spatio-temporal and 12 gait) extracted from IMUs. They propose a novel nonlinear model to obtain the level of gait task (UPDRS) but from the sensor features, which accurately rates 84.9% of the subjects. This result is better than the other ML algorithms tested (SVM, NB, Linear Regression), although the dataset is from a small sample of 25 PwPD and 28 HC.

By last, a paper based on a public dataset, where Khamparia et al. (2024) investigate whether it is possible to distinguish typical walking patterns from a gait brain disorder. The Daphnet Freezing of Gait Dataset (Daniel Roggen, 2010) is of medium size when compared to the previous literature discussed, it has 9 attributes and 237 rows, referring to sensor data of the IMU obtained during rehabilitation. Although this article mentions several objectives, the most relevant one is to evaluate the performance of the seven ML models experimented (LogReg, SVM, RF, DT, kNN, NB and Perceptron). Among these models, some dimensionality reduction techniques were applied to evaluate their impact on the models' results, with kNN and RF being the ones that achieved higher metrics (accuracy, precision, recall, F1-score).

4. Methodology

4.1. Data access

The data used for this project is considered secondary data as it was not collected for the purpose of this study. The data was provided by a healthcare institution that stores patient assessment data relating to neuropsychology, nursing, nutrition, occupational therapy, physiotherapy and speech therapy.

This project starts with one dataset containing anonymised data that is compliant with the General Data Protection Regulation (GDPR), as the patients have signed an informed consent form allowing their anonymised data to be used for research purposes. Approval was also required from the NOVA IMS Ethics Committee, details of which are available in Appendix A.

4.2. Tools

Python was chosen as the programming language (*Python*, 2024) to work with this data, using the Visual Studio Code Integrated Development Environment (*Microsoft*, 2025), which facilitates the use of several libraries. These include NumPy (Harris et al., 2020; *NumPy*, 2025), pandas (McKinney, 2010; *Pandas*, 2024) and Matplotlib (Hunter, 2007; *Matplotlib*, 2025) which are dedicated to data exploration and visualisation, and Scikit-Learn (Pedregosa et al., 2011; *Scikit Learn*, 2025) was used to implement and evaluate the models. Microsoft Excel (*Microsoft Corporation*, 2025) was also used to handle some of the intermediate files in either .csv or .xlsx format. To deal with the feature importance, SHAP values (*SHAP*, 2025), will be used to explain some of the outputs.

Regarding AI tools, GitHub Copilot (2025) was used as an assistant tool for code development, applied with moderation and cognizance for any ethical concerns. To assist with the writing process, including translating words and rewriting some sentences, the free versions of DeepL Translate and DeepL Write were used (*DeepL*, 2025).

4.3. Cross-Industry Standard Process

CRISP-DM is a well-known methodological framework and was described in Chapter 2. In this section, each stage is contextualised within the scope of this project. It should be noted that this process is iterative, as seen earlier on Figure 1.

4.3.1. BUSINESS UNDERSTANDING

As previously referred in *2.2. Health, disease and rehabilitation concepts*, PD is a complex disease to diagnose, also because the clinical criteria used by the neurologists has some subjective parameters (the interpretation of the evaluation scales tend to be dependent on each practitioner experience). In most cases, the distinction between PD and AP is made based on the rapid progression of the symptoms and how the patient responds to the usual medications. However, the earlier this distinction is detected, the better the outcome for the patient, as more effective therapeutic and rehabilitation interventions can be designed.

Therefore, the main objective of this thesis is to search for an ML approach, a binary classification model, that can distinguish between PwPD and PwAP, based on rehabilitation (physiotherapy, neuropsychology and others) assessments. At a later stage, after obtaining a robust model, the aim is to identify which of the features contribute the most to this classification. The data used in this project originates from a private dataset from routine assessments performed at a medical institution.

4.3.2. DATA UNDERSTANDING

The data available for this project was provided on a comma-separated values (.csv) file. This dataset (named DPF) started with 44 917 records from 3 763 patients and contains 5 471 features. These features contain demographic information (age, gender, date of assessment), clinical information (diagnosis, date of diagnosis, records of other exams, pharmacologic therapy) and every possible answer to each one of the questionnaires from all areas of the facility (nursing, speech therapy, physiotherapy, etc.). Many of these last features (possible answers) were extracted in a similar way of a one-hot encoding which explains the high number of features (but with extremely sparse information). Before any further exploration on the variables, 4 301 duplicated records were removed.

FEATURES

Since there was no feature dictionary available to interpret and understand all the variables, the first approach to the selection was to get a features list. Although there was no dictionary, the institution provided a concise list of the protocol assessments done for both AP and PD patients, from which the keywords of Table 2 were extracted. Then, based on the knowledge acquired during the literature review and the keywords related to the context of this project, the list of features was screened. This selection task consisted in reading each of the variables label and try to understand its meaning (sometimes it was difficult to determine whether the answer belonged to the same questionnaire or was the start of a different one; other times, erratic characters made the text illegible). Then, if it was related to the project or reported in the literature (such as age, gender, area of assessment, number of hospitalizations) it was

included. Another consideration was to select features that contained any of the keywords mentioned on Table 2 (even if the remaining text was not perceptible). After all this lengthy process, 93 features remained (and 5 378 were excluded).

Table 2 - List of keywords used to select the variables

Keywords list
patient; cgi; clinical global impression; pharmacological therapies; esas; edmonton symptom assessment system; mds-updrs; mmse; mini mental state examination; moca; montreal cognitive assessment; n-fog; new freezing of gait; nine-hole peg test; pgi; patient global impression; hy; hoehn and yahr; se; schwab and england; who; world health organization; whodas; who disability assessment schedule

TARGET

Regarding the target feature "diagnosis", it was another complex variable to work with. In the records, it is possible to find cases as the following examples: 'diagnosis_'; 'diagnosis_dd-mm-yyyy'; 'diagnosis1_, diagnosis2_mm-yyyy'; 'diagnosis_yyyy'; some diagnoses are in Portuguese, others in English and even for the same diagnosis there are different words. The first step on these inconsistencies was to keep only the records with diagnosis of interest, by selecting the rows that contained any of the keywords: [*Parkinson*, *atypical*, *corticobasal*, *Lewy*, *system*, *supranuclear*]. This step removed 17 465 rows and left 23 151. Then on another reading exploration of the dataset, it was found too many cases of *Vascular Parkinsonism* that is not a diagnosis of interest. So, all the rows that had the word *vascular* were removed, accepting that it might be dropping some complex cases of interest out of the 1594 rows removed.

Since the reality is not as clear as one patient exclusively having PD or AP. Some patients have other concomitant diagnoses (such as a previous stroke, dementia, neoplasms, etc.), the diagnosis of these patients will be designated as *Parkinsons_cx* or *AP_cx* (cx: complex diagnosis). Other patients have been first diagnosed with PD and later with AP, due to its sudden rapid progression or other evaluation criteria. These cases will be excluded since the date of each diagnose is not always mentioned, and the date of the corresponding assessment is also not identifiable. However, it is remarked the relevance of this corrected data for future research on prognosis. The records where the diagnosis was only PD or AP were kept as *Parkinsons_b* or *AP_b* (b: basic).

The final step to deal with the target inconsistency, was to export the dataset into an Excel file, order the diagnoses alphabetically, and rename them in batches. This step allowed to verify the quality of the selection and identify some details to solve. There were more records deleted: 262 because the search by *Parkinson* also identified different kinds of parkinsonism outside the scope of the project, and 484 had both PD and AP diagnosis.

This raw dataset demanded an exhaustive data understanding, which was crucial for a proper contextualisation into the data, and to prepare it for the following developments. When reaching the end of this step, the date of birth feature was replaced by age (to avoid concerns about identifiable patterns on the anonymous data), which led to the loss of 262 records due to missing values on date of birth.

4.3.3. DATA PREPARATION

This subchapter details the iterative process of the steps between exploratory data analysis and preprocessing in this project, including the reasoning behind the decisions made.

Starting from the handled dataset of the previous subchapter, which had 20 549 records and 93 features. The decision to drop the records missing a date of birth, was based on the relevance of the age feature. Although there are well-known and reliable ways for handling missing values, this project aims to avoid introducing any underlying bias when doing so.

About the target feature ("diagnosis"), with the changes done until now, there are 882 unique patients, that translate into the following number of records for each diagnosis: 14 152 (*Parkinsons_b*); 5 546 (*AP_b*); 535 (*Parkinsons_cx*); 316 (*AP_cx*). Even though the dataset is not yet ready to use, it is possible to anticipate that it is unbalanced and there will be a need to use techniques to mitigate the target imbalance and to wisely choose evaluation metrics that consider this imbalance.

FEATURES

To perform the exploration of the 93 features, they were divided into 13 categories: Global, CGI, PGI, PhT, ESAS, mini-BEST, MMSE, MoCA, N-FoG, 9HPT, MDS-UPDRS, S&E, WHO. In the light of the amounts of missing data found, it is worth remembering that these features were initially selected based solely on their description and potential relevance to the project.

Starting with 15 Global features: "Pt_ID", "Age", "Diagnosis", "Area", "Battery", "Assessment", "Date", "Period", "Applicable", "#_hospitalization", "Score", "Question", "Aborted", "Gender", "HY_score".

The patient code ("Pt_ID") which is an anonymized and random number, allows to keep track of how many unique patients are there, it does not have any missing values, and will be dropped later. The "Age" feature has no missing values and ranges from 1 to 106 years old as seen on Figure 2. Considering that the diagnosis of PD in patients under 50 years is considered as EOPD (a diagnosis with different characteristics) and that the infants may be due to an incorrect date of birth, these 184 records were dropped. Then, it was used the Interquartile Range (IQR) to determine the bounds for the outliers, which were 55.5 and 99.5 years old, which led to more 179 rows being deleted. There were 20 186 records left.

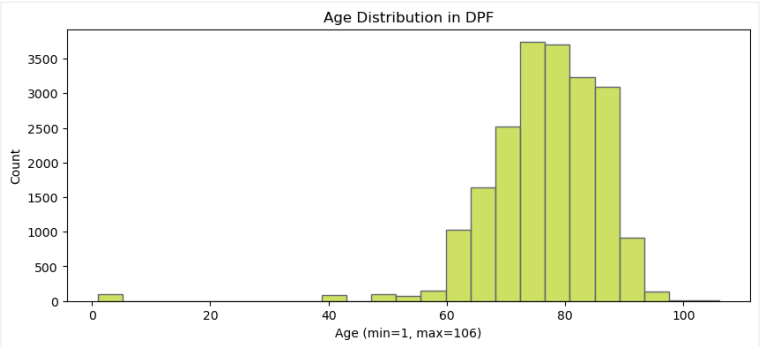

Figure 2 - Feature "Age" distribution on DPF dataset.

The “Diagnosis” was considered inside the Global category, however, due to it being the target variable, its analysis was repeated in the end of the process.

As for the “Area”, its values refer to which sector has done the assessment, and it has 14 785 values (marked as ‘missing’). “Battery” feature has a similar issue, with 14 960 (‘missing’) values and the remaining ones refer to the sector and condition (‘admission’ or ‘discharge’). The “Assessment” feature is the one to keep for now, as it has more information than the previous two (provides the name and type of questionnaire or evaluation) and has no missing values nor values listed as ‘missing’.

Referring to the “Date”, “Period”, and “Applicable” features, they do not have missing values but have more than 50% of its values filled as ‘aborted’, ‘missing’, invalid date’, ‘na’ or ‘__’. A similar situation when exploring “#_hospitalization” where 73% of its values are ‘__’ which can either be a missing value or the absence of hospitalization. “Score” has 5327 missing values (‘NaN’) and only have values of ‘na’ or ‘aborted’. “Question” also has 73% of ‘__’ values, and the remaining values are difficult to relate to the project. Finally, the features “Aborted”, “Gender” and “HY_score” only present inconclusive values of ‘0’, ‘__’, ‘true’, ‘True’, ‘false’, ‘False’ and missing values (‘NaN’).

Based on the Global features, it was possible to observe cases with several records for the same patient, on the same date, on the same battery, but with a different assessment and subsequently different values in the following features. However, due to incongruencies in the date formats (including missing values), it was decided that instead of merging these observations (which potentially had incorrect dates), each record would be used as a unique patient. To better support this decision, it is important to remember that the duration in time is a major point in the decision between AP and PD and would be beneficial to have a model that investigates other critical features. Besides, a patient can have the first appointment and assessment records in any time-point of either the diseases (PD or AP), so the unknow diversity of disease progression could make the classification algorithm to have a broader applicability.

Continuing with the CGI category, out of 8 features, 5 were dropped because all the values were '__', 'missing', 'mam' or 'na'. The 3 features that remained have some filled scores (on the correct scale from 1-5), although most of the values are '__'. However, they might be worth keeping, if a merging approach of the rows comes to be tested later. The PGI features have a quite similar exploration to the CGI ones. From 14 initial features, only 3 have been kept and the reasoning was the same. As for the PhT features, none of the pharmacological therapies variables was kept, because they only contained '0', '__', 'mam', and 'missing'.

Regarding the ESAS, 5 out of 6 features had too many values missing (the same ones as represented before) and only 1 feature was kept. The same happened in mini-BEST, MMSE, MoCA and N-FoG grouped features, were each category had 3 features, and only 1 per category has remained. On the 9HPT there were 3 features, and none was selected due to the missing values.

The MDS-UPDRS category has the 21 features with the most potential, since 14 of them were selected to be kept on the dataset. The remaining were dropped due to values missing, non-relevant data or the data being too unbalanced to be handled. On S&E features only 1 feature was selected, and 6 features were dropped for the same reasons as before. Finally, on WHO category, none of the 4 features had enough valid data to work with.

After this exploration, 60 features were dropped, 33 remain and are ready to a deeper level of exploration and preprocessing. At this step, an .xlsx file was exported to open on Excel and visualize the data coherence of the current DPF dataset.

PREPROCESSING

Although there have been previous steps that are considered as preprocessing, this data has needed such a great amount of iterative work around it, that this approach of splitting the steps may be easier to interpret or reproduce.

On the observation of the Excel file, it was easily noticed that a considerable number of rows, from the feature "Date" onwards, were exclusively filled with '__', mainly the ones where "Date" had the 'aborted' value. On other features there were many 'mam' and 'na'. And 8 of the MDS-UPDRS features were duplicated, so they were dropped, along with 1 PGI and 1 CGI features with the same characteristics.

Before any action on these findings, DPF had 20 186 records and 33 features. But then, when the features following the "Date" were again explored, there was a need to drop the rows were the values for the full row (after "Date") were exclusively filled with any combination of the values 'NaN', '0', 'mam', '__', 'na'. These actions had a major impact on the DPF size, that has reduced to 1 822 records and 23 features.

Then, 4 features ("Area", "Battery", "Assessment" and "Date") were dropped, as they were no longer needed to check for data coherence. Exploring again the remaining missing values, the 4 features that have more than 80% of *NaN* ("CGI_c", "CGI_s", "PGI_c", "PGI_s") had to be dropped. At this point, when visualizing the Excel, there was still plenty of *NaN* values. On the remaining 15 features, 12 of them correspond to assessment features from which 6 belong to the MDS-UPDRS category. Considering the reported relevance of this category, it was determined that rows that have more than 5 *NaN* values in those 12 variables should be deleted. When re-working on the missing values, only the features with less than 20% of missing values were selected to be kept on DPF ("ESAS_pd", "MMSE", "mini-BT", "MoCA", "N-FoG", "SE" have been dropped). Since the dataset was quickly reducing its size, the need to use methods to input missing values started to arise.

Considering this almost ready DPF dataset, currently with 1 580 records of 521 unique patients and 9 features, it is time to assess the number of records an remaining unique patients of the target feature "Diagnosis" on Table 3.

Table 3 - Target feature "Diagnosis" values.

Diagnoses	Number of records	Unique patients
Parkinson_b	1040	323
AP_b	467	165
Parkinson_cx	43	20
AP_cx	30	13

The decision to exclude the complex cases is supported by the small sample of records that would not be enough to develop a well-functioning model, and merging these cases could lead to a more difficult classification on the models. However, this approach is left as a suggestion to future research.

As the final changes to the dataset: 1 119 records were dropped, because when searching in the 6 remaining MDS-UPDRS features, these rows were filled with '0'; and the anonymized and random "Pt_ID" feature was deleted.

FINAL DATASET AND SPLIT

After all the exploration and processing, the final DPF dataset contains 388 records and 8 variables, where the target has been encoded to *AP_b* = 0; *Parkinson_b* = 1. The dataset was exported as a .csv file, to be used in the next steps in a different notebook (focused on models). DPF will be split inside a loop, where each split will account for 75% of the data for training and 25% to the test set, while keeping the same proportions of each of the target classes.

4.3.4. MODELLING

The models to be experimented for classification between PD and AP are: Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF), Multi-layer Perceptron (MLP), Support Vector Classification (SVC) and Extreme Gradient Boosting (XGB). These algorithms will be applied to the DPF dataset, and the same experiment design will be repeated on DPFss dataset which is the DPF dataset but scaled and balanced with synthetic data.

The experiment design consisted of evaluating these 6 models and compare their performances throughout 30 independent runs. To ensure the variability across runs, each one used a different random seed applied to the whole pipeline. This means that all the steps of the same run use the same random seed: stratified train-test split, cross-validations (cv) and even the models itself (the ones that have *random_state* as a parameter).

On Appendix B, the code from the modelling process is presented. To support a robust comparison and ensure the randomness in the data splitting, the process was repeated 30 times using different random seeds (without replacement). The outer loop of splitting the data, guarantees that all models will be trained and evaluated on the same data partitions per run. The variability is introduced across runs with the random seed, since in each run, the data was split in 75% for train and 25% for the test set, while stratified by the target to keep the same class proportions.

Appendix C demonstrates the code used to train each model. The dedicated function *run_model* encapsulates the hyperparameter optimization, the training, and the scoring strategy of train and test. *GridSearchCV* (GS) was run on the train set, to perform the hyperparameter tuning, using the *StratifiedKFold* (SkF) as the 5-fold cv splitting strategy. The parameter grids for each model are available to consult on Appendix D. Regarding the scoring strategy, it was multi-objective: F1-Score Macro (F1M); Balanced Accuracy (BAcc); Recall Macro and F1-Score Weighted (F1W). However, the model selection was based on F1M.

To each of the best models given by the GS (on each run), their mean scores (from 5-fold SkF) of train and test, the best parameters and the runtime were stored in structured *DataFrame* (df) and exported to .csv and .xlsx formats. To avoid the expense of more time and computational power, the 180 trained models were saved (6 models x 30 runs), allowing for a future deployment or interpretability analysis.

To perform the statistical analysis, it was compiled the following: 6 arrays (1 per model) containing the test values of F1M; 6 arrays with the BAcc test values; a df summarizing the performance metrics across all 30 runs using the mean and standard deviation (std).

After identifying the best performing model through the statistical testing and to fulfil the second objective of identifying the features' impact, Shapley Additive Explanations (SHAP) will be put into practice. SHAP values need to use the data to perform, so the previously saved trained model will not be enough. By gathering the information on the best model and its

parameters, run number and random state, it is possible to replicate the data split, and train the model over again. Then it is possible to run SHAP on the test set and wait for the chosen visualization (*beeswarm*) to appear. Its interpretation will be addressed in the next chapter.

Regarding the "second" dataset DPFss, the experiment design is basically the same as previously explained for DPF and available on Appendix B and Appendix D. Some few modifications were made to the *run_model* function on Appendix E, since the pipeline function had to be adapted to account for the scaling (*StandardScaler*) and SMOTE (Synthetic Minority Oversampling Technique). This decision on a second dataset with synthetic data, was due the amount of data that was lost during the processing. It may help to understand how the models behave on this synthetic data.

4.3.5. EVALUATION

The chosen metrics for the evaluation of the models are macro F1-score (F1M) and balanced accuracy (BAcc). The F1M is preferred than the F1W, because the macro component gives an equal weight to each class. The BAcc considers for the specificity and sensitivity, while in this problem it represents the same as the recall-macro (accounting the same for both classes even though one has much less cases than the other). In this clinical context, neither of the classes (PD or AP) have more importance than the other, because this is a distinction between two diseases with similar prognostic value. Since there is no problem of false positive vs false negative, nor exists a class of healthy controls, the F1M better evaluates the models' performance. However, similar studies report by F1W, without clarifying the reasoning or any justification for their selection.

As detailed in the previous subchapter, the evaluation of the results goes in layers, being the first evaluation inside the GS, where the 5-fold cv using *SkF* is run with the *refit='f1_macro'* parameter, to optimize using the F1M. So, this GS will determine a *best_estimator* that will be used to predict on the test set. The second layer is that each model will be run 30 times, each time with a different random seed, applied from the split of the data. This code composition was motivated by the concept of nested-cv.

On the statistical tests, the Friedman test was applied to evaluate whether the differences in the models' performance were statistically significant. If necessary, this can be followed by a *post-hoc* analysis to determine where the differences lie between the models being compared. In this thesis context, the further analysis can be done by Wilcoxon tests with Bonferroni or Holm corrections, or by Nemenyi tests. These statistical tests account for the variance of the splits and ensure that the answer given to the first research question is evidence-based.

4.3.6. DEPLOYMENT

Deployment is the final step in the CRISP-DM framework. This stage considers a well-defined algorithm, with all the clinical trials approved and statistical evidence of its work. The model would then be deployed for use by certified healthcare professionals (the specific professional category depends on the legislation in each country) in clinical settings as an auxiliary diagnostic tool. As this stage goes beyond the scope of this thesis, it will not be explored any further.

5. Results and Discussion

5.1. Binary classification task – DPF dataset

One of the objectives of this thesis is to identify a ML model that best classifies if a patient suffers from PD or AP, given the rehabilitation assessments data. The evaluation of the models was based on the performance metrics of F1M and BAcc, which both consider for the class imbalance of the minority class (AP).

The mean performance metrics over 30 independent runs, for each of the experimented models is presented on Table 4 . It contains the mean scores for both the training and the test sets, accompanied by the std.

Table 4 - Performance metrics on DPF (mean ± std)

Model	Train_F1M	Train_BAcc	Test_F1M	Test_BAcc
LR	0.5417 ± 0.0213	0.6477 ± 0.0241	0.5228 ± 0.0221	0.6234 ± 0.0451
NB	0.5050 ± 0.0301	0.5158 ± 0.0242	0.4964 ± 0.0358	0.5076 ± 0.0254
RF	0.6365 ± 0.0266	0.6345 ± 0.0313	0.5601 ± 0.0770	0.5664 ± 0.0752
MLP	0.6087 ± 0.0216	0.6029 ± 0.0241	0.5443 ± 0.0552	0.5440 ± 0.0497
SVC	0.5804 ± 0.0241	0.6366 ± 0.0333	0.5484 ± 0.0656	0.5960 ± 0.0826
XGB	0.5892 ± 0.0402	0.5835 ± 0.0344	0.6051 ± 0.0810	0.5958 ± 0.0681

From this Table 4 scores, and evaluating on the test metrics, it seems that XGB has the best performance, when considering for the F1M. If focusing on BAcc, then LR seems to perform better. Table 5 presents a rank of the models, sorted by the test scores (F1M, BAcc), what gives a better perception on the models' performance when accounting for different metrics.

Table 5 - Models ranked by metrics (mean test scores on DPF)

Rank	F1M		BAcc	
1	XGB	0.6051	LR	0.6234
2	RF	0.5601	SVC	0.5960
3	SVC	0.5484	XGB	0.5958
4	MLP	0.5443	RF	0.5664
5	LR	0.5228	MLP	0.5440
6	NB	0.4964	NB	0.5076

F1-SCORE MACRO

Starting with the statistical analysis, based on the test scores of F1M, the first evaluation was the Friedman test. This is a non-parametric test that allows to detect the differences between multiple ML models across multiple test attempts, in this case across 30 runs.

All the tests on F1M were run assuming a significance threshold of α=0.05, and the null hypothesis (H_0) is that there is no significant difference in performance between the models (i.e. the six models perform equally).

The result of the Friedman test is p-value=1.034×10^{-8}, which means H_0 is rejected, so there is a significant difference between the models' performance.

As there is a significant difference in at least one of the models, a pairwise comparison is required. In this project context, there are 3 common *post-hoc* analyses to be performed after the Friedman test: Wilcoxon tests with Bonferroni or Holm corrections and Nemenyi tests. The selection of one of these tests is dependent on what the researcher aims to. In this case, it was decided to experiment the 3 approaches and compare its results. The approach to the 3 analyses is similar, a short loop function, that calculates the p-value for each pairwise combination, and a correction factor to be applied to the p-value or to the alpha.

The results of the Wilcoxon test with Bonferroni correction were, assuming a corrected α=0.0033, that there is a significant difference in the following 7 pairwise combinations: (LR, XGB); (NB, RF); (NB, MLP); (NB, SVC); (NB, XGB); (MLP, XGB); (SVC, XGB).

The results of the Wilcoxon test with Holm correction (done after getting all the uncorrected p-values) were the following 9 combinations: (LR, NB); (LR, XGB); (NB, RF); (NB, MLP); (NB, SVC); (NB, XGB), (RF, XGB); (MLP, XGB); (SVC, XGB).

Finally, the Nemenyi test results were that there is a significant difference between the following 7 pairwise combinations: (LR, XGB); (NB, RF); (NB, MLP); (NB, SVC); (NB, XGB); (MLP, XGB); (SVC, XGB).

The results were compiled on Table 6, while resorting to Table 5 as a brief reminder of the rank between the models, according to F1M. The reading of this table can be done as: the first column ('Combination') corresponds to all the 15 pairwise combinations; the second column ('WB') corresponds to the Wilcoxon test with Bonferroni correction; the third column ('WH') is the Wilcoxon test with Holm correction; and the fourth column ('N') refers to the Nemenyi test. Regarding the rows: if a cell has '*ns*' value, it means that the test failed to reject the null so there is no significant difference; if it has the abbreviation of any of the models, it means the test rejected the null hypothesis (i.e. there is a significant difference between the F1M of the two pairwise models) and when using Table 5 to check the rank, the highest model 'wins' so it gets the abbreviation on the cell. The rows marked in bold text refer to the combinations of the 3 highest ranked models.

To conclude the evaluation reasoning, by taking the 3 higher ranked models (F1M – XGB, RF, SVC), it is possible to check that:

- the 1st (XGB) and the 2nd (RF) ranked do not have a significant difference (WB, N tests), but it has a significant difference on the WH test;

- there is no significant difference between the 2nd (RF) and the 3rd (SVC) in any of the tests;

- there is a significant difference between the 1st (XGB) and the 3rd (SVC) in all the tests.

So, since XGB and RF do not have a significant difference in 2 out of 3 tests, it is possible to affirm that, based on F1M, there are 2 equally good models at predicting PD vs AP cases.

Table 6 - Pairwise comparisons (F1M - DPF)

Combination	**WB**	**WH**	**N**
LR vs NB	ns	LR	ns
LR vs RF	ns	ns	ns
LR vs MLP	ns	ns	ns
LR vs SVC	ns	ns	ns
LR vs XGB	XGB	XGB	XGB
NB vs RF	RF	RF	RF
NB vs MLP	MLP	MLP	MLP
NB vs SVC	SVC	SVC	SVC
NB vs XGB	XGB	XGB	XGB
RF vs MLP	ns	ns	ns
RF vs SVC	**ns**	**ns**	**ns**
RF vs XGB	**ns**	**XGB**	**ns**
MLP vs SVC	ns	ns	ns
MLP vs XGB	XGB	XGB	XGB
SVC vs XGB	**XGB**	**XGB**	**XGB**

BALANCED ACCURACY

Given that there are only two main evaluation metrics, it was decided to also perform a statistic analysis based on the BAcc, using the same methodology as for F1M.

All the tests on BAcc were run assuming a significance threshold of α=0.05, and the null hypothesis (H_0) is that there is no significant difference in performance (i.e. the models perform equally). The result of the Friedman test is p-value=1.179×10^{-10}, which means H_0 is rejected, so there is a significant difference between the models' performance.

The results of the Wilcoxon test with Bonferroni correction (α=0.0033): there is a significant difference in the following 8 pairwise combinations: (LR, NB); (LR, RF); (LR, MLP); (NB, RF); (NB, MLP); (NB, SVC); (NB, XGB); (MLP, XGB).

The results of the Wilcoxon test with Holm correction (after getting all the uncorrected p-values) were the following 8 combinations: (LR, NB); (LR, RF); (LR, MLP); (NB, RF); (NB, MLP); (NB, SVC); (NB, XGB); (MLP, XGB).

Finally, the Nemenyi test results were that there is a significant difference between the following 8 pairwise combinations: (LR, NB); (LR, RF); (LR, MLP); (NB, RF); (NB, SVC); (NB, XGB); (MLP, SVC); (MLP, XGB).

The results were arranged on Table 7, with the help of Table 5 to check the rank and scores between the models, according to BAcc. The interpretation of this table was explained in the previous subchapter (F1-score macro).

To conclude the evaluation of the models, based on BAcc, when selecting the 3 higher ranked models (BAcc – LR, SVC, XGB), it is observed that, in all tests, there is no significant difference on the pairwise comparison of these models. When checking for the 4th ranked model (RF) and comparing to LR, it has a significant difference. So, it is possible to declare that, based on BAcc, there are 3 equally performing models at the proposed binary classification task.

Table 7 - Pairwise combinations (test BAcc - DPF)

Combination	**WB**	**WH**	**N**
LR vs NB	LR	LR	ns
LR vs RF	LR	LR	LR
LR vs MLP	LR	LR	LR
LR vs SVC	**ns**	**ns**	**ns**
LR vs XGB	**ns**	**ns**	**ns**
NB vs RF	RF	RF	RF
NB vs MLP	MLP	MLP	ns
NB vs SVC	SVC	SVC	SVC
NB vs XGB	XGB	XGB	XGB
RF vs MLP	ns	ns	ns
RF vs SVC	ns	ns	ns
RF vs XGB	ns	ns	ns
MLP vs SVC	ns	ns	SVC
MLP vs XGB	XGB	XGB	XGB
SVC vs XGB	**ns**	**ns**	**ns**

5.2. SHAP VALUES - DPF

Considering the statistical test results, based on F1M there are 2 equally performing models (XGB and RF); while based on BAcc there are 3 models that do not show any pairwise significant difference (LR, SVC, XGB).

To perform the SHAP values, only one model per metric will be chosen. Since they do not present any significant difference, there is less relevance on why to choose one or the other. In this case, for F1M it will be XGB; and for BAcc the LR.

To better interpret the SHAP plots, it is relevant to remember that the X-axis (SHAP-value) represents the impact of each feature on the model output. So, values to the left suggest a push towards predicting the negative class (AP=0), while values to the right suggest a push toward the positive class (PD=1). As for the Y-axis, the list of features is ranked based on their overall importance (mean absolute SHAP value), from most (top) to least (bottom) influential. Finally, each dot represents a single SHAP value for one instance and one feature, and the colour depicts the feature value (higher values are red/pink, lower values are blue).

Figure 3 displays the plot of the XGB model. The possible interpretation is: "mds_updrs_pIII" is the most impactful contributor to predictions. The feature higher values shift predictions to the (negative class, AP=0). So, it suggests that worse motor scores reduce the likelihood of being predicted as PD. The "Age" feature has some impact as well, but mixed, with a high distribution on the SHAP values, but disperse in the features values. "mds_updrs" shows mixed influence, and it is caused because it's the sum of all the mds_updrs parts. The remaining features have smaller SHAP values (more centred around 0).

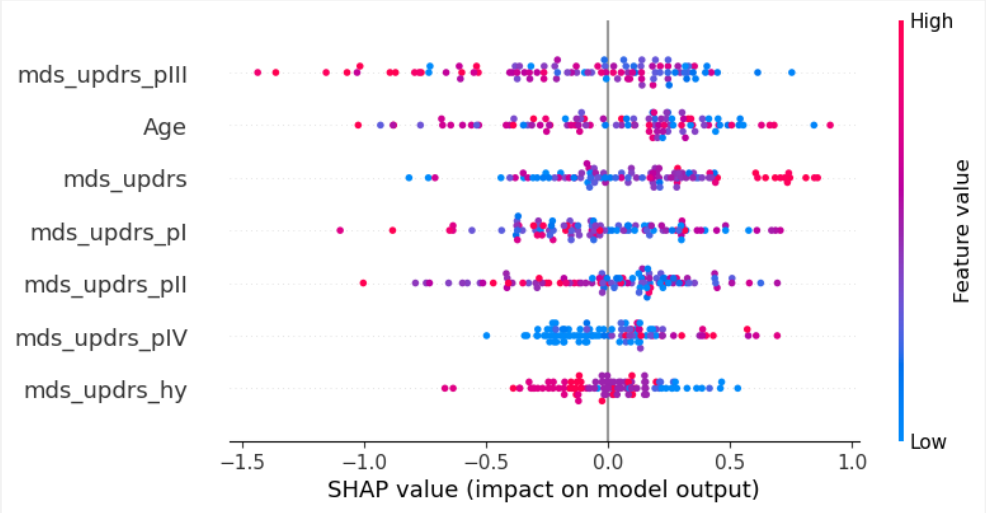

Figure 3 - SHAP values plot (*beeswarm*) of XGB (DPF)

The best performing LR model (by BAcc), produced the following SHAP-values, plotted on Figure 4. The possible interpretation of this plot is that "mds_updrs_III" has the largest overall impact on predictions, and the higher feature values tend to push the prediction to AP=0. "mds_updrs" also is highly influential, but it is contradictory to the first feature, due to the weight of the next feature "mds_updrs_pIV" as it has a high quantity of lower value points. Considering that this happens on both evaluations, it would be useful to test just with the independent parts of mds-updrs. The feature "Age" has a clear distribution, as its higher values tend to tend to push the prediction to the left (AP=0), lower ages contribute to PD, and older ages to AP, which can be explained by the temporal factor that determines the progression of the disease. "mds_updrs_pII" also has a clear distribution, but a lower impact (rank). The remaining features are less influential because their SHAP values are close to zero, meaning they contribute minimally to the model's predictions.

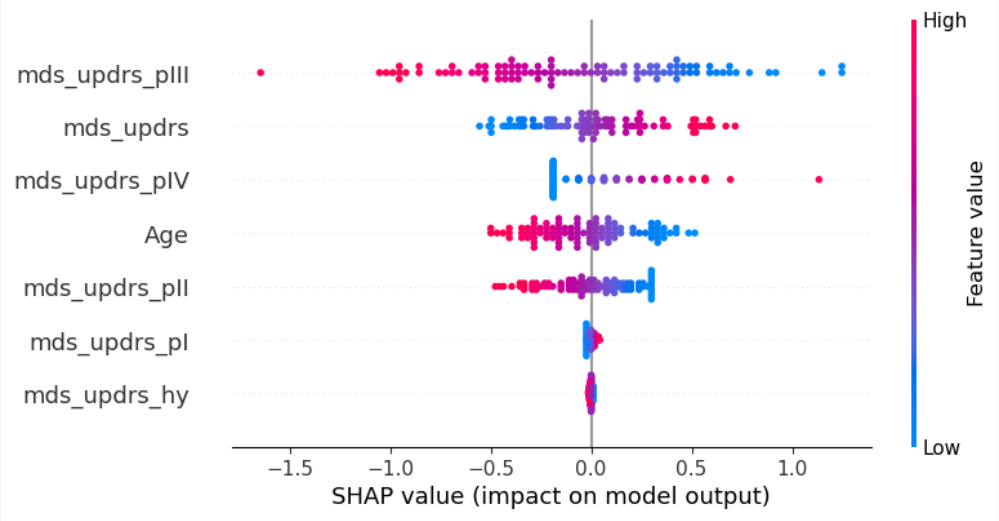

Figure 4 - SHAP values plot (beeswarm) of LR (DPF)

5.3. BINARY CLASSIFICATION TASK – DPFSS DATASET

Due to the abrupt difference from the original dataset to the processed one, it surged the need to use other methods to deal with the data. The creation of a second dataset works as a comparison to the approach taken on the raw dataset, as this one includes synthetic data and a scaling process.

Over 30 independent runs, the mean scores of the performance metrics were saved, for each of the tested models. This information is reflected on Table 8, that contains the scores (mean±std) for both the train and test sets.

Table 8 - Performance metrics on DPFss (mean ± std)

Model	Train_F1M	Train_BAcc	Test_F1M	Test_BAcc
LR	0.5424 ± 0.0191	0.6438 ± 0.0275	0.5281 ± 0.0407	0.6292 ± 0.0594
NB	0.5099 ± 0.0260	0.6145 ± 0.0339	0.5182 ± 0.0383	0.6444 ± 0.0524
RF	0.6320 ± 0.0190	0.6547 ± 0.0246	0.5892 ± 0.0409	0.6065 ± 0.0457
MLP	0.6253 ± 0.0295	0.6405 ± 0.0290	0.5956 ± 0.0635	0.6102 ± 0.0699
SVC	0.5990 ± 0.0256	0.6381 ± 0.0349	0.5510 ± 0.0557	0.5950 ± 0.0767
XGB	0.5927 ± 0.0250	0.6004 ± 0.0267	0.5997 ± 0.0552	0.6019 ± 0.0565

In this Table 8, it is possible to have an overview and create some expectations about the results of the statistical tests. On Table 9, it is observable that depending on the preferred metric, the ranking of the models goes almost to the opposite. Table 9 presents a rank of the models, sorted by the test scores (F1M, BAcc), what gives a better perception on the performance of the models when accounting for different metrics.

Table 9 - Models ranked by metrics (mean test scores on DPFss)

Rank	F1M		BAcc	
1	XGB	0.5997	NB	0.6444
2	MLP	0.5956	LR	0.6292
3	RF	0.5892	MLP	0.6102
4	SVC	0.5510	RF	0.6065
5	LR	0.5281	XGB	0.6019
6	NB	0.5182	SVC	0.5950

F1-SCORE MACRO

Starting with the analysis, based on F1M test scores, the first evaluation is the Friedman test, which got a p-value of 9.216×10^{-11}. Considering that all the statistical tests on were run assuming a significance threshold of $\alpha=0.05$, and the null hypothesis (H_0) is that there is no significant difference in performance. So, for the first test, the null hypothesis is rejected, which means that the models have a significant difference in performance.

The approach will be the same as before, to perform 3 *post-hoc* analysis. For the Wilcoxon test with Bonferroni correction, the corrected alpha will be the same (0.0033), as it depends on the number of combinations, and keep the same 15 ones. The results of this test are that there is a significant difference in 8 comparisons: (LR, RF); (LR, MLP); (LR, XGB); (NB, RF); (NB, MLP); (NB, XGB); (MLP, SVC); (SVC, XGB).

The Wilcoxon test with Holm correction resulted in 9 pairwise comparisons that are significantly different: (LR, RF); (LR, MLP); (LR, XGB); (NB, RF); (NB, MLP); (NB, SVC); (NB, XGB); (MLP, SVC); (SVC, XGB). To end, the results of Nemenyi test were 7 combinations that present a significant difference: (LR, RF); (LR, MLP); (LR, XGB); (NB, RF); (NB, MLP); (NB, SVC); (NB, XGB). These pairwise comparison results have been compiled in Table 10, while consulting Table 9 to get the ranks of each model, in this step according to F1M.

To conclude the evaluation of the models, based on F1M, when selecting the 3 higher ranked models (F1M – XGB, MLP, RF), it is observed that, in all tests, there is no significant difference on the pairwise comparison of these models. When checking for the 4th ranked model (SVC) and comparing to XGB, it has a significant difference. So, it is possible to declare that, based on F1M, there are 3 equally performing models at the proposed binary classification task.

Table 10 - Pairwise comparisons (F1M - DPFss)

Combination	**WB**	**WH**	**N**
LR vs NB	ns	ns	ns
LR vs RF	RF	RF	RF
LR vs MLP	MLP	MLP	MLP
LR vs SVC	ns	ns	ns
LR vs XGB	XGB	XGB	XGB
NB vs RF	RF	RF	RF
NB vs MLP	MLP	MLP	MLP
NB vs SVC	SVC	SVC	SVC
NB vs XGB	XGB	XGB	XGB
RF vs MLP	**ns**	**ns**	**ns**
RF vs SVC	ns	ns	ns
RF vs XGB	**ns**	**ns**	**ns**
MLP vs SVC	MLP	MLP	ns
MLP vs XGB	**ns**	**ns**	**ns**
SVC vs XGB	XGB	XGB	ns

BALANCED ACCURACY

To replicate the same approach of DPF, the Friedman test will be applied as a statistical analysis of the models, to detect differences between its performances. Based on BAcc scores of DPFss, the result of the Friedman test is a p-value=0.1069, which is bigger than the significance level. In this particular case, it fails to reject the null hypothesis, which means that the models do not have any significant differences between them. So, there is no need for any *post-hoc* analysis.

6. Conclusions and Future Research

6.1. Conclusions

The growth in Artificial Intelligence (AI) innovations has been exponential across all sectors, including healthcare. Its impact is undeniable, especially when focused on improving patient outcomes and readjust the workload of healthcare professionals. The speed of evolution has only been slowed down by the rigorous laws and ethical concerns that characterize healthcare. With this focus, the endeavour of this thesis was to contribute towards the development an auxiliary diagnostic tool.

The increase in life expectancy over the past years goes along with an increment in the prevalence of Parkinson's Disease (PD). The economic burden of PD has been rising, with 11 million cases reported in 2021, and this figure is set to rise further as the disease progresses. Atypical Parkinsonism (AP) shares common symptoms with PD, but differentiating between the two diagnoses has a significant impact on the progression of each disease and, consequently, affects the quality of life of patients.

Machine Learning (ML) has been applied to a diversity of complex problems surrounding PD, including with diverse types of data. However, it was found a gap in the literature regarding the distinction between PD and AP using ML models. Since none of the diseases has a cure, the management of the disease progression takes a bigger relevance in the patient outcomes.

The aim of this thesis was to develop a model with sufficient performance in a binary classification task of distinguishing between PD and AP that, in the long term, could potentially become an auxiliary diagnostic tool for healthcare professionals. The earlier the diagnosis, the greater the benefits for the patient, as it becomes possible to target and contain disease progression more effectively. The other objective was to identify what features have more weight on the models' classification, which allows for a better understanding of the model and might lead to new research on those variables.

In this study, a rigorous experimental design was implemented to evaluate and compare, across 30 independent runs, the performance of six ML classifiers: Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF), Multi-layer Perceptron (MLP), Support Vector Classifier (SVC), and XGBoost (XGB). Each run used a different random seed to generate a stratified 75/25 train-test split, ensuring variability across runs. Within each run, models were trained and fine-tuned using 5-fold cross-validation (*StratifiedKFold*) and *GridSearchCV* to identify optimal hyperparameters. This repeated evaluation strategy provided a robust foundation for assessing model stability, generalization, and statistical significance of performance differences.

Although the raw dataset was of a considerable size, after all the processing it turned out as a small dataset, with few variables. To account for the imbalance data and try to improve the models scores, a second dataset was created and tested by passing the original one inside a *Pipeline* with *StandardScaler* and SMOTE.

In conclusion of this thesis, it is possible to answer to both proposed research objective (RO) as follows:

RO1 – To identify a ML model which best classifies whether a patient has PD or APD.

Considering the F1-score macro (F1M) on the DPF dataset, XGB and RF are the two best performing models. The XGB has a mean F1M of 0.6051 ± 0.0810, and the RF has a mean of 0.5601 ± 0.0770. The statistical testing showed that there is no significance difference between the performances of these models, which can be transposed to its classification ability.

If considering the balanced accuracy (BAcc) on DPF, then three models show no significant differences between their performance metrics, what can be converted into their classification ability. The three models and its BAcc scores are: (LR, 0.6234 ± 0.0451); (SVC, 0.5960 ± 0.0826) and (XGB, 0.5958 ± 0.0681).

Although these were not the expected positive results, they are statistically reliable. The code implementation of this methodology allows for further research with new features and more data. The discussion on negative results is as important as on positive ones, because it allows to understand what does not work, and to experiment new and different approaches on future studies.

RO2 – Identify which features have the most influence on the previous classification.

These conclusions are limited to the small number of features, but according to the SHAP values, the feature that have the most impact on the distinction between AP and PD is the “mds_updrs_pIII” because it is ranked first on both plots and have a higher distribution of values. When considering the best F1M XGB model, the feature “Age” is the second most relevant although its dispersion of the features values. If considering the best BAcc LR model, the second most relevant feature is “mds_updrs”, with a clear distribution. Since this feature is the sum of the four parts of MDS-UPDRS, and one of the parts (feature “mds_updrs_pII) is exactly the opposite, it might be cancelling some understanding on this feature.

6.2. LIMITATIONS AND FUTURE RESEARCH

Regarding a final overview of this project, some of the limitations have been described throughout the thesis. In summary, each of the supported decisions made, during the data curating stage, might have a significant impact on the performance of the models. Data quality also proved to be a serious limitation, since it required extensive manual processing beyond the standard frameworks, and was scarcer than anticipated. The conservative approach of avoiding introducing noise or bias to the data (as it is medical data) in exchange for fewer records (due to the missing values) has a repercussion on the results. The synthetic data from one of the datasets was used primarily for comparison purposes rather than to improve the model. Another limitation is the performance metrics selected to evaluate the models. Which, in this case with the low performance it would not be necessary, but with a high performing model it could make sense to compare the results against medical experts.

There are many suggestions for future research as the results were not so positive as expected, but first it is important to remember the importance of a multidisciplinary team when the projects involve many different specialties. To this project it is suggested to repeat the approach but include more variables, and assuming proper techniques to fill the missing values. As for future research on the topic, the duration of the disease could be incorporated by having at least three evaluations of specific time-points, aiming to add a prognostic value to the research. Other approaches could be experimented, such as restraining the data to each of the evaluation sectors (physiotherapy, nursing, etc.) or focusing on the distinction of the AP subcategories. The ideas of experimenting with other models or gathering more data to explore the distinction in patients with concomitant pathologies are also suggested.

As a final reflection, any future research aimed at improving human life holds immense value. Progress in science is not solely measured by positive outcomes; negative or inconclusive results also play a meaningful role to the collective understanding, guiding future efforts with greater clarity. As we continue to innovate and explore new frontiers, it is imperative to do so with a strong commitment to ethical responsibility, transparency, and equity. Only by upholding these principles can we ensure that scientific advancement genuinely serves society and contributes to the responsible transformation of it.

Bibliographical References

Benavoli, A., Corani, G., & Mangili, F. (2016). Should we really use post-hoc tests based on mean-ranks? *J. Mach. Learn. Res.*, *17*(1), 152–161.

Bologna, M., Espay, A. J., Fasano, A., Paparella, G., Hallett, M., & Berardelli, A. (2023). Redefining Bradykinesia. *Movement Disorders : Official Journal of the Movement Disorder Society*, *38*(4), 551–557. https://doi.org/10.1002/mds.29362

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The Balanced Accuracy and Its Posterior Distribution. *2010 20th International Conference on Pattern Recognition*, 3121–3124. https://doi.org/10.1109/ICPR.2010.764

Cardoso, F., Goetz, C. G., Mestre, T. A., Sampaio, C., Adler, C. H., Berg, D., Bloem, B. R., Burn, D. J., Fitts, M. S., Gasser, T., Klein, C., de Tijssen, M. A. J., Lang, A. E., Lim, S.-Y., Litvan, I., Meissner, W. G., Mollenhauer, B., Okubadejo, N., Okun, M. S., … Trenkwalder, C. (2024). A Statement of the MDS on Biological Definition, Staging, and Classification of Parkinson's Disease. *Movement Disorders*, *39*(2), 259–266. https://doi.org/10.1002/mds.29683

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 1–27. https://doi.org/10.1145/1961189.1961199

Chaudhuri, K. R., Azulay, J.-P., Odin, P., Lindvall, S., Domingos, J., Alobaidi, A., Kandukuri, P. L., Chaudhari, V. S., Parra, J. C., Yamazaki, T., Oddsdottir, J., Wright, J., & Martinez-Martin, P. (2024). Economic Burden of Parkinson's Disease: A Multinational, Real-World, Cost-of-Illness Study. *Drugs - Real World Outcomes*, *11*(1), 1–11. https://doi.org/10.1007/s40801-023-00410-1

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. 785–794. https://doi.org/10.1145/2939672.2939785

Daniel Roggen, M. P. (2010). *Daphnet Freezing of Gait* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C56K78

Dara, S., Dhamercherla, S., Jadav, S. S., Babu, C. M., & Ahsan, M. J. (2022). Machine Learning in Drug Discovery: A Review. *Artificial Intelligence Review*, *55*(3), 1947–1999. https://doi.org/10.1007/s10462-021-10058-4

DeepL. (2025). https://www.deepl.com/en/write

Derby Hospitals NHS Foundation Trust. (2003). *Lindop Parkinson's Assessment Scale*. Derby Hospitals NHS Foundation Trust. https://www.parkinsons.org.uk/sites/default/files/2017-12/lindopparkinsonsassessmentscale.pdf

Deutschländer, A. B., Ross, O. A., Dickson, D. W., & Wszolek, Z. K. (2018). Atypical parkinsonian syndromes: A general neurologist's perspective. *European Journal of Neurology*, *25*(1), 41–58. https://doi.org/10.1111/ene.13412

Devarajan, J. P., Sreedharan, V. R., & Narayanamurthy, G. (2023). Decision Making in Health Care Diagnosis: Evidence From Parkinson's Disease Via Hybrid Machine Learning. *IEEE Transactions on Engineering Management*, *70*(8), 2719–2731. IEEE Transactions on Engineering Management. https://doi.org/10.1109/TEM.2021.3096862

Diallo, R., Edalo, C., & Awe, O. O. (2025). Machine Learning Evaluation of Imbalanced Health Data: A Comparative Analysis of Balanced Accuracy, MCC, and F1 Score. In O. O. Awe & E. A. Vance (Eds.), *Practical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network, USA* (pp. 283–312). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72215-8_12

Excel. (2025). [Computer software]. Microsoft Corporation. https://www.microsoft.com/en-us/microsoft-365/excel

eXtreme Gradient Boosting. (2025). *XGBoost Python Package*. XGBoost Python Package — Xgboost 3.0.2 Documentation. https://xgboost.readthedocs.io/en/stable/python/index.html

Facciorusso, S., Spina, S., Reebye, R., Turolla, A., Calabrò, R. S., Fiore, P., & Santamato, A. (2023). Sensor-Based Rehabilitation in Neurological Diseases: A Bibliometric Analysis of Research Trends. *Brain Sciences*, *13*(5), Article 5. https://doi.org/10.3390/brainsci13050724

García, S., & Herrera, F. (2008). An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. *Journal of Machine Learning Research*, *9*, 2677–2694.

GitHub Copilot. (2025). GitHub. https://github.com/features/copilot

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M. B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A. E., Lees, A., Leurgans, S., LeWitt, P. A., Nyenhuis, D., … LaPelle, N. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Movement Disorders*, *23*(15), 2129–2170. https://doi.org/10.1002/mds.22340

Grażyńska, A., Urbaś, W., Antoniuk, S., Adamczewska, K., Bień, M., Chmiela, T., & Siuda, J. (2020). Comparative analysis of non-motor symptoms in patients with Parkinson's Disease and atypical parkinsonisms. *Clinical Neurology and Neurosurgery*, *197*, 106088. https://doi.org/10.1016/j.clineuro.2020.106088

Gupta, R., Kumari, S., Senapati, A., Ambasta, R. K., & Kumar, P. (2023). New era of artificial intelligence and machine learning-based detection, diagnosis, and therapeutics in Parkinson's disease. *Ageing Research Reviews*, *90*, 102013. https://doi.org/10.1016/j.arr.2023.102013

Han, Y., Liu, X., Zhang, N., Zhang, X., Zhang, B., Wang, S., Liu, T., & Yi, J. (2023). Automatic Assessments of Parkinsonian Gait with Wearable Sensors for Human Assistive Systems. *Sensors*, *23*(4), Article 4. https://doi.org/10.3390/s23042104

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hinojosa Lee, M. C., Braet, J., & Springael, J. (2024). Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores. *Applied Sciences*, *14*(21), Article 21. https://doi.org/10.3390/app14219863

Hirani, R., Noruzi, K., Khuram, H., Hussaini, A. S., Aifuwa, E. I., Ely, K. E., Lewis, J. M., Gabr, A. E., Smiley, A., Tiwari, R. K., & Etienne, M. (2024). Artificial Intelligence and Healthcare: A Journey through History, Present Innovations, and Future Possibilities. *Life*, *14*(5), Article 5. https://doi.org/10.3390/life14050557

Höglinger, G. U., Kassubek, J., Csoti, I., Ehret, R., Herbst, H., Wellach, I., Winkler, J., & Jost, W. H. (2017). Differentiation of atypical Parkinson syndromes. *Journal of Neural Transmission*, *124*(8), 997–1004. https://doi.org/10.1007/s00702-017-1700-4

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jacob, D., Aubonnet, R., Recenti, M., Audardottir, S. A., Ida Ivarsdottir, T., Burgunder, B., Escalona, I. M. I., Colacino, A., Bjornsdottir, A., Petersen, H., & Gargiulo, P. (2022). Assessing Early-stage Parkinson's Disease Using BioVRSea. *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*, 271–276. https://doi.org/10.1109/MetroXRAINE54828.2022.9967502

Jankovic, J. (2008). Parkinson's disease: Clinical features and diagnosis. *Journal of Neurology, Neurosurgery & Psychiatry*, *79*(4), 368–376. https://doi.org/10.1136/jnnp.2007.131045

Kamran, I., Naz, S., Razzak, I., & Imran, M. (2021). Handwriting dynamics assessment using deep neural network for early identification of Parkinson's disease. *Future Generation Computer Systems*, *117*, 234–244. https://doi.org/10.1016/j.future.2020.11.020

Keus, S., Munneke, M., Graziano, M., Paltamaa, J., Pelosin, E., Domingos, J., Ramaswamy, B., Prins, J., Struiksma, C., Rochester, L., Nieuwboer, A., & Bloem, B. (2014). European Physiotherapy Guideline for Parkinson's Disease. *KNGF/ParkinsonNet*. https://www.parkinsonnet.nl/app/uploads/sites/3/2019/11/eu_guideline_parkinson_guideline_for_pt_s1.pdf

Khamparia, A., Gupta, D., Maashi, M., & Mengash, H. A. (2024). Cognitive driven gait freezing phase detection and classification for neuro-rehabilitated patients using machine learning algorithms. *Journal of Neuroscience Methods*, *409*, 110183. https://doi.org/10.1016/j.jneumeth.2024.110183

Khan, M., & Sherani, A. M. K. (2025). Leveraging AI for Efficient Healthcare Workforce Management: Addressing Staffing Shortages and Reducing Burnout. *Global Journal of Computer Sciences and Artificial Intelligence*, *1*(1), Article 1. https://doi.org/10.70445/gjcsai.1.1.2025.43-54

Krones, F., Marikkar, U., Parsons, G., Szmul, A., & Mahdi, A. (2025). Review of multimodal machine learning approaches in healthcare. *Information Fusion*, *114*, 102690. https://doi.org/10.1016/j.inffus.2024.102690

Kruse, R., Mostaghim, S., Borgelt, C., Braune, C., & Steinbrecher, M. (2022). Multi-layer Perceptrons. In R. Kruse, S. Mostaghim, C. Borgelt, C. Braune, & M. Steinbrecher (Eds.), *Computational Intelligence: A Methodological Introduction* (pp. 53–124). Springer International Publishing. https://doi.org/10.1007/978-3-030-42227-1_5

Lee, T. K., & Yankee, E. L. (2022). A review on Parkinson's disease treatment. *Neuroimmunology and Neuroinflammation*, *8*, 222. https://doi.org/10.20517/2347-8659.2020.58

Li, Y., McLernon, D. J., Counsell, C. E., & Macleod, A. D. (2024). Incidence and risk factors of institutionalisation in Parkinson's disease and atypical parkinsonism. *Parkinsonism & Related Disorders*, *118*, 105928. https://doi.org/10.1016/j.parkreldis.2023.105928

Liu, J., & Xu, Y. (2022). T-Friedman Test: A New Statistical Test for Multiple Comparison with an Adjustable Conservativeness Measure. *International Journal of Computational Intelligence Systems*, *15*(1), 29. https://doi.org/10.1007/s44196-022-00083-8

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777.

Martinez-Martin, P., Skorvanek, M., Rojo-Abuin, J. M., Gregova, Z., Stebbins, Glenn. T., Goetz, C. G., & Group, members of the Q. S. (2018). Validation study of the hoehn and yahr scale included in the MDS-UPDRS. *Movement Disorders*, *33*(4), 651–652. https://doi.org/10.1002/mds.27242

Matloff, N. S. (2017). *Statistical regression and classification: From linear models to machine learning*. CRC Press. https://doi.org/10.1201/9781315119588

Matplotlib (Version 3.9.2). (2025). [Python library]. Matplotlib. https://matplotlib.org/stable/

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the Python in Science Conference*, 56–61. https://doi.org/10.25080/majora-92bf1922-00a

Mehanna, R., Smilowska, K., Fleisher, J., Post, B., Hatano, T., Pimentel Piemonte, M. E., Kumar, K. R., McConvey, V., Zhang, B., Tan, E.-K., Savica, R., & the International Parkinson and Movement Disorder Society Task Force on Early Onset Parkinson's Disease. (2022). Age Cutoff for Early-Onset Parkinson's Disease: Recommendations from the International

Parkinson and Movement Disorder Society Task Force on Early Onset Parkinson's Disease. *Movement Disorders Clinical Practice*, *9*(7), 869–878. https://doi.org/10.1002/mdc3.13523

Mitchell, T. M. . (1997). *Machine Learning* (Nachdr.). McGraw-Hill. https://www.cs.cmu.edu/~tom/files/MachineLearningTomMitchell.pdf

Mohsin Khan, M., Shah, N., Shaikh, N., Thabet, A., Alrabayah, T., & Belkhair, S. (2025). Towards secure and trusted AI in healthcare: A systematic review of emerging innovations and ethical challenges. *International Journal of Medical Informatics*, *195*, 105780. https://doi.org/10.1016/j.ijmedinf.2024.105780

Munhoz, R. P., Tumas, V., Pedroso, J. L., & Silveira-Moriyama, L. (2024). The clinical diagnosis of Parkinson's disease. *Arquivos de Neuro-Psiquiatria*, *82*, 1–10. https://doi.org/10.1055/s-0043-1777775

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. *Journal of the American Geriatrics Society*, *53*(4), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

NumPy (Version 1.26.4). (2025). [Python library]. NumPy. https://numpy.org/doc/stable/user/index.html

Pandas (Version 2.2.3). (2024). [Python library]. pandas. https://pandas.pydata.org/pandas-docs/version/2.2.3/

Pedregosa, F., Pedregosa, F., Varoquaux, G., Varoquaux, G., Org, N., Gramfort, A., Gramfort, A., Michel, V., Michel, V., Fr, L., Thirion, B., Thirion, B., Grisel, O., Grisel, O., Blondel, M., Prettenhofer, P., Prettenhofer, P., Weiss, R., Dubourg, V., … Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12 (85)*, 2825–2830. https://jmlr.csail.mit.edu/papers/volume12/pedregosa11a/pedregosa11a.pdf

Pereira, D. G., Afonso, A., & Medeiros, F. M. (2015). Overview of Friedman's Test and Post-hoc Analysis. *Communications in Statistics - Simulation and Computation*, *44*(10), 2636–2653. https://doi.org/10.1080/03610918.2014.931971

Pogrzeba, L., Neumann, T., Wacker, M., & Jung, B. (2019). Analysis and Quantification of Repetitive Motion in Long-Term Rehabilitation. *IEEE Journal of Biomedical and Health*

Informatics, *23*(3), 1075–1085. IEEE Journal of Biomedical and Health Informatics. https://doi.org/10.1109/JBHI.2018.2848103

Postuma, R. B., Berg, D., Stern, M., Poewe, W., Olanow, C. W., Oertel, W., Obeso, J., Marek, K., Litvan, I., Lang, A. E., Halliday, G., Goetz, C. G., Gasser, T., Dubois, B., Chan, P., Bloem, B. R., Adler, C. H., & Deuschl, G. (2015). MDS clinical diagnostic criteria for Parkinson's disease. *Movement Disorders*, *30*(12), 1591–1601. https://doi.org/10.1002/mds.26424

Python 3.12.7 documentation (Version 3.12.7). (2024). [Programming language]. Python Software Foundation. https://docs.python.org/release/3.12.7/

Russo, M., Amboni, M., Volzone, A., Cuoco, S., Camicioli, R., Di Filippo, F., Barone, P., Romano, M., Amato, F., & Ricciardi, C. (2024). Kinematic and Kinetic Gait Features Associated With Mild Cognitive Impairment in Parkinson's Disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *32*, 2676–2687. https://doi.org/10.1109/TNSRE.2024.3431234

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Scikit learn (Version 1.5.1). (2025). [Machine learning library]. scikit-learn. https://scikit-learn/stable/user_guide.html

scikit-learn: bacc. (2025). *Balanced_accuracy_score*. Scikit-Learn: Balanced Accuracy Score. https://scikit-learn/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html

scikit-learn: f1. (2025). *F1_score*. Scikit-Learn: F1-Score. https://scikit-learn/stable/modules/generated/sklearn.metrics.f1_score.html

scikit-learn: LR. (2025). *Logistic Regression*. Scikit-Learn: LogisticRegression. https://scikit-learn/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

scikit-learn: MLP. (2025). *Multi-layer Perceptron Classifier*. Scikit-Learn: MLPClassifier. https://scikit-learn/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

scikit-learn: NB. (2025). *Naïve Bayes*. Scikit-Learn: Naive Bayes. https://scikit-learn/stable/modules/naive_bayes.html

scikit-learn: RF. (2025). *Random Forest Classifier*. Scikit-Learn: RandomForestClassifier. https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

scikit-learn: SCV. (2025). *Support Vector Classification*. Scikit-Learn: SVC. https://scikit-learn/stable/modules/generated/sklearn.svm.SVC.html

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

SHAP (Version v0.48.0). (2025). [Computer software]. SHAP (SHapley Additive exPlanations). https://shap.readthedocs.io/en/latest/

Smith, M. D., Brazier, D. E., & Henderson, E. J. (2021). Current Perspectives on the Assessment and Management of Gait Disorders in Parkinson's Disease. *Neuropsychiatric Disease and Treatment*, *Volume 17*, 2965–2985. https://doi.org/10.2147/NDT.S304567

Templeton, J. M., Poellabauer, C., & Schneider, S. (2022). Classification of Parkinson's disease and its stages using machine learning. *Scientific Reports*, *12*(1), 14036. https://doi.org/10.1038/s41598-022-18015-z

Thomas, M., Lenka, A., & Kumar Pal, P. (2017). Handwriting Analysis in Parkinson's Disease: Current Status and Future Directions. *Movement Disorders Clinical Practice*, *4*(6), 806–818. https://doi.org/10.1002/mdc3.12552

Tombaugh, T. N., & McIntyre, N. J. (1992). The Mini-Mental State Examination: A Comprehensive Review. *Journal of the American Geriatrics Society*, *40*(9), 922–935. https://doi.org/10.1111/j.1532-5415.1992.tb01992.x

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Tsakanikas, V., Ntanis, A., Rigas, G., Androutsos, C., Boucharas, D., Tachos, N., Skaramagkas, V., Chatzaki, C., Kefalopoulou, Z., Tsiknakis, M., & Fotiadis, D. (2023). Evaluating Gait Impairment in Parkinson's Disease from Instrumented Insole and IMU Sensor Data. *Sensors*, *23*(8), Article 8. https://doi.org/10.3390/s23083902

University of Washington. (2024). *Institute for Health Metrics and Evaluation (IHME). GBD Results.* https://vizhub.healthdata.org/gbd-results/

Virameteekul, S., Revesz, T., Jaunmuktane, Z., Warner, T. T., & De Pablo-Fernández, E. (2023). Clinical Diagnostic Accuracy of Parkinson's Disease: Where Do We Stand? *Movement Disorders*, *38*(4), 558–566. https://doi.org/10.1002/mds.29317

Visanji, N. P., Madan, P., Lacoste, A. M. B., Buleje, I., Han, Y., Spangler, S., Kalia, L. V., Hensley Alford, S., & Marras, C. (2021). Using artificial intelligence to identify anti-hypertensives as possible disease modifying agents in Parkinson's disease. *Pharmacoepidemiology and Drug Safety*, *30*(2), 201–209. https://doi.org/10.1002/pds.5176

Visual Studio Code (Version 1.101). (2025). [IDE]. Microsoft Corporation. https://code.visualstudio.com/docs

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, *1*, 29–39. https://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

Zhang, H. (2004). The Optimality of Naive Bayes. *American Association for Artificial Intelligence*. https://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf

Zhang, J., Fan, Y., Liang, H., & Zhang, Y. (2024). Global, regional and national temporal trends in Parkinson's disease incidence, disability-adjusted life year rates in middle-aged and older adults: A cross-national inequality analysis and Bayesian age-period-cohort analysis based on the global burden of disease 2021. *Neurological Sciences*. https://doi.org/10.1007/s10072-024-07941-7

Appendix A

ETHICS COMMITTEE OF NOVA IMS APPROVAL

Outlook

RE: NOVA IMS | Ethics Committee - NEED REVIEW

De Ethics Committee <ethicscommittee@novaims.unl.pt>
Data sex, 13/06/2025 14:43
Para Susana Cristina Norberto Pires <20230540@novaims.unl.pt>; Leonardo Vanneschi <lvanneschi@novaims.unl.pt>
Cc Ethics Committee <ethicscommittee@novaims.unl.pt>

Dear Susana Pires,
Dear Professor Leonardo Vanneschi,

Thank you for completing the Research Ethics Checklist. Concerning the use of secondary anonymized health data provided by the institution, please note that ethical research conduct and institutional regulations require clear documentation regarding data access and usage conditions.
Although the dataset has already been anonymized prior to your access, and its collection was covered by informed consent procedures signed by patients — which explicitly allow the use of their data for research purposes — it remains essential to secure formal confirmation from the data-providing institution. This confirmation should verify:

- That the data was collected in the course of routine services and subsequently anonymized.
- That informed consent was obtained from patients, allowing for the anonymized data to be used in academic or scientific research.
- That your access is limited strictly to anonymized records, with no possibility of re-identification.
- That the institution authorizes the use of this anonymized dataset for your specific academic project.

This formal statement ensures compliance with data protection standards, preserves the rights of data subjects, and provides documentation of the institution's authorization and oversight.
Provided this written confirmation is obtained, your study may proceed. As the data is fully anonymized and informed consent for research use has been embedded into the institution's processes, no major ethical risks have been identified.

Project No.: **DSCI2025-6-139134**

Project Title: **Parkinson's Disease and Atypical Parkinsonism in the context of rehabilitation: a machine learning approach**

Principal Researcher: **Susana C. N. Pires**

according to the regulations of the Ethics Committee of NOVA IMS and MagIC Research Center this project was considered to meet the requirements of the NOVA IMS Internal Review Board, being considered **APPROVED** on 13/06/2025.

It is the Principal Researcher's responsibility to ensure that all researchers and stakeholders associated with this project are aware of the conditions of approval and which documents have been approved.

The Principal Researcher is required to notify the Ethics Committee, via amendment or progress report, of

- Any significant change to the project and the reason for that change;
- Any unforeseen events or unexpected developments that merit notification;
- The inability of the Principal Researcher to continue in that role or any other change in research personnel involved in the project.

Lisbon, 13/06/2025
NOVA IMS Ethics Committee
ethicscommittee@novaims.unl.pt

This email serves as formal proof of ethical approval. If required for inclusion in a thesis, dissertation, or any other academic documentation, a PDF version of this message may be created and attached accordingly.

Cristina Oliveira
Gestora executiva do centro de investigação MagIC | *Executive manager of the Information Management Research Center (MagIC)*
Find out more about our research at https://magic.novaims.unl.pt/en/
Team member of RM Roadmap - Co-creating the future of Research Management (https://rmroadmap.eu/)
https://orcid.org/0000-0002-0887-7961

Appendix B

EXPERIMENTAL DESIGN

```
results_all = []
random_thirty = np.random.choice(range(0, 1000), size=30, replace=False)

# Count the total time of all runs
st_full_time = time.perf_counter()

# Run 30 times with different random states
for idx, seed in enumerate(random_thirty):
    print(f"Run {idx+1}/30 - random_state={seed}")

    # Split the data with the same random state of the run
    # Each model is trained on the same data split (will have different splits across different runs)
    X_tval, X_test, y_tval, y_test = train_test_split(X, y, test_size=0.25, random_state=seed, stratify=y)

    # Scaling and SMOTE - inside the run_model function (pipeline)

    # Start run timer (of each run)
    st_run_time = time.perf_counter()

    # LR
    result_lr = run_model(model_name="LR", estimator=LogisticRegression(random_state=seed),
                          param_grid=prm_grid_LR, scoring=scores_gs,
                          X_tval=X_tval, y_tval=y_tval,
                          X_test=X_test, y_test=y_test, seed=seed)
    results_all.append(result_lr)

    # NB
    result_nb = run_model("NB", GaussianNB(), prm_grid_NB, scores_gs, X_tval, y_tval, X_test, y_test, seed)
    results_all.append(result_nb)

    # RF
    results_rf = run_model("RF", RandomForestClassifier(random_state=seed), prm_grid_RF, scores_gs, X_tval, y_tval, X_test, y_test, seed)
    results_all.append(results_rf)

    # MLP
    results_mlp = run_model("MLP", MLPClassifier(random_state=seed), prm_grid_MLP, scores_gs, X_tval, y_tval, X_test, y_test, seed)
    results_all.append(results_mlp)

    # SVC
    results_svc = run_model("SVC", SVC(random_state=seed), prm_grid_SVC, scores_gs, X_tval, y_tval, X_test, y_test, seed)
    results_all.append(results_svc)

    # XGBoost
    results_xgb = run_model("XGB", xgb.XGBClassifier(random_state=seed), prm_grid_XGB, scores_gs, X_tval, y_tval, X_test, y_test, seed)
    results_all.append(results_xgb)

    # End run timer
    run_time = time.perf_counter() - st_run_time
    print(f"Run {idx+1} completed in {run_time:.2f} seconds ({run_time/60:.2f} minutes) \n")


full_time = time.perf_counter() - st_full_time
print(f"Full 30 runs completed in {full_time/60:.2f} minutes ({full_time/3600:.2f} hours)")
```

Appendix C

DPF *RUN_MODEL* FUNCTION

```python
# Define the function to run each model
def run_model(model_name, estimator, param_grid, scoring, X_tval, y_tval, X_test, y_test, seed):

    # Define the skf for the gs-cv (in this step all models will get the same random_state, different per run)
    skf_cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=seed)

    # Define the pipeline with the estimator (will be useful to scale the data)
    pipeline = Pipeline([('clf', estimator)])

    # Prepare the GridSearch for the model
    gs = GridSearchCV(estimator=pipeline,
                      param_grid=param_grid,
                      cv=skf_cv,
                      scoring=scoring,
                      refit='f1_macro',
                      n_jobs=-1)

    # Fit the GS on 75% of data (Split outside this function - but inside the same random_state per run)
    t_start = time.perf_counter()
    gs.fit(X_tval, y_tval)
    elapsed = time.perf_counter() - t_start

    # Get the results and best index (f1_macro)
    results = gs.cv_results_
    best_idx = gs.best_index_

    # Save the train results
    train_scores = {'f1_macro': results['mean_test_f1_macro'][best_idx],
                    'balanced_accuracy': results['mean_test_balanced_accuracy'][best_idx],
                    'recall_macro': results['mean_test_recall_macro'][best_idx]}

    # Predict on test set using the best estimator (GS) per run
    y_pred = gs.best_estimator_.predict(X_test)

    # Save the test results
    test_scores = {'f1_macro': f1_score(y_test, y_pred, average='macro'),
                   'balanced_accuracy': balanced_accuracy_score(y_test, y_pred),
                   'recall_macro': recall_score(y_test, y_pred, average='macro')}

    return {'model': model_name,
            'train_scores': train_scores,
            'test_scores': test_scores,
            'fit_time_sec': elapsed,
            'best_idx': best_idx,
            'best_score': gs.best_score_,       # mean f1_macro from gs-cv scores
            'best_params': gs.best_params_,
            'best_estimator': gs.best_estimator_}
```

Appendix D

GridSearchCV parameter grids

```
# Logistic Regression
prm_grid_LR = {'clf__C': [0.001, 0.01, 0.1, 1, 10],
               'clf__penalty': ['l2', 'none', 'l1', 'elasticnet'],
               'clf__class_weight': [None, 'balanced'],
               'clf__solver': ['lbfgs', 'liblinear', 'sag', 'saga'],
               'clf__max_iter': [100, 500, 1000]}

# Naive Bayes
prm_grid_NB = {'clf__var_smoothing': [1e-9, 1e-8, 1e-7, 1e-6, 1e-5, 1e-4, 1e-3]}

# Random Forest
prm_grid_RF = {'clf__n_estimators': [100, 10, 25, 50],
               'clf__criterion': ['gini', 'entropy', 'log_loss'],
               'clf__max_depth': [None, 5, 10],
               'clf__min_samples_split': [2, 5, 10],
               'clf__min_samples_leaf': [1, 2, 4],
               'clf__max_features': ['sqrt', 'log2', None],
               'clf__bootstrap': [True, False],
               'clf__class_weight': [None, 'balanced']}

# Multi-layer Perceptron
prm_grid_MLP = {'clf__hidden_layer_sizes': [(100,), (50,), (75, 20), (50, 25)],
                'clf__activation': ['relu', 'identity', 'logistic', 'tanh'],
                'clf__solver': ['adam', 'lbfgs', 'sgd'],
                'clf__alpha': [0.0001, 0.001, 0.01],
                'clf__learning_rate': ['constant', 'invscaling', 'adaptive'],
                'clf__max_iter': [100, 250, 500],
                'clf__shuffle': [True, False],
                'clf__tol': [0.0001, 0.001, 0.01],
                'clf__early_stopping': [True]}

# Support Vector Classification
prm_grid_SVC = {'clf__C': [0.001, 0.01, 0.1, 1, 10],
                'clf__kernel': ['rbf', 'linear', 'poly', 'sigmoid'],
                'clf__degree': [1, 2, 3, 4],
                'clf__gamma': ['scale', 'auto'],
                'clf__class_weight': [None, 'balanced'],
                'clf__max_iter': [100, 500, 1000]}

# XGBoost
prm_grid_XGB = {'clf__objective': ['binary:logistic', 'binary:hinge']}
```

Appendix E

DPFSS *RUN_MODEL* FUNCTION

```
# Define the function to run each model
def run_model(model_name, estimator, param_grid, scoring, X_tval, y_tval, X_test, y_test, seed):

    # Define the skf for the gs-cv (in this step all models will get the same random_state, different per run)
    skf_cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=seed)

    # Define the pipeline with the estimator (1st scaler then smote)
    pipeline = imbPipeline([('scaler', StandardScaler()), ('smote', SMOTE(random_state=seed)), ('clf', estimator)])

    # Prepare the GridSearch for the model
    gs = GridSearchCV(estimator=pipeline,
                      param_grid=param_grid,
                      cv=skf_cv,
                      scoring=scoring,
                      refit='f1_macro',
                      n_jobs=-1)

    # Fit the GS on 75% of data (Split outside this function - but inside the same random_state per run)
    t_start = time.perf_counter()
    gs.fit(X_tval, y_tval)
    elapsed = time.perf_counter() - t_start

    # Get the results and best index (f1_macro)
    results = gs.cv_results_
    best_idx = gs.best_index_

    # Save the train results
    train_scores = {'f1_macro': results['mean_test_f1_macro'][best_idx],
                    'balanced_accuracy': results['mean_test_balanced_accuracy'][best_idx],
                    'recall_macro': results['mean_test_recall_macro'][best_idx]}

    # Predict on test set using the best estimator (GS) per run
    y_pred = gs.best_estimator_.predict(X_test)

    # Save the test results
    test_scores = {'f1_macro': f1_score(y_test, y_pred, average='macro'),
                   'balanced_accuracy': balanced_accuracy_score(y_test, y_pred),
                   'recall_macro': recall_score(y_test, y_pred, average='macro')}

    return {'model': model_name,
            'train_scores': train_scores,
            'test_scores': test_scores,
            'fit_time_sec': elapsed,
            'best_idx': best_idx,
            'best_score': gs.best_score_,       # f1_macro from gs-cv scores
            'best_params': gs.best_params_,
            'best_estimator': gs.best_estimator_}
```

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa