

**UNDERSTANDING OF PARKINSON'S DISEASE THROUGH EXPLAINABLE
ARTIFICIAL INTELLIGENCE APPLIED TO SPEECH MEASUREMENTS**

A DISSERTATION SUBMITTED TO MANCHESTER METROPOLITAN UNIVERSITY
FOR THE DEGREE OF MASTER OF SCIENCE
IN THE FACULTY OF SCIENCE AND ENGINEERING



2024

By
Ramses Moreno De La Cruz
Department of Computing and Mathematics

Contents

Abstract	vii
Declaration.....	viii
Acknowledgements.....	ix
Abbreviations	x
CHAPTER 1. - INTRODUCTION	4
1.1 Project Overview	4
1.2 Aim and Objectives.....	5
1.2 Dissertation Project Structure	5
CHAPTER 2. - LITERATURE REVIEW.....	7
2.1 Introduction.....	7
2.2 Parkison's disease.....	7
2.3 Voice disorders.....	10
2.4 Perceptual Analysis and Acoustic Analysis.....	11
2.5 Survey of Machine and Deep Learning in Acoustic Analysis	12
CHAPTER 3. - DATA	14
3.1 Introduction.....	14
3.2 Key insights about the dataset	14
3.3 Description of the Dependent and Independent variables	16
CHAPTER 4. - DESIGN, IMPLEMENTATION AND RESULTS	21
4.1 Introduction.....	21
4.2 Set up the Environment / Loading the dataset	21

4.3	Splitting the whole dataset	22
4.4	Exploratory Data Analysis.....	23
4.5	Feature Engineering	25
4.6	Contrasts of the Boosted Trees model and the Convolutional Neural Network	25
4.7	Boosted Tree Model	26
4.8	Convolutional Neural Network.....	28
4.9	Boosted Trees / Convolutional Neural Network - Model Interpretability and Explainability.....	35
4.10	Boosted Tree Model – Feature Selection.....	39
4.11	Boosted Tree Model (Improved Version) – Training Phase, Validation Phase and Test Phase	43
4.12	Boosted Tree Model (Improved Version) – Analysis and Evaluation of the results	44
4.13	Boosted Tree Model (Improved Version - Model Interpretability and Explainability.....	45
4.14	Key Insights of the assessment and evaluation of the results	46
	CHAPTER 5. - FURTHER WORK.....	49
5.1	Introduction.....	49
5.2	Chapter 3. Data – Challenges	49
5.3	Chapter 4. Data – Design, Implementation and Results - Challenges....	49
5.4	Future ideas.....	50
	CHAPTER 6. - CONCLUSIONS	52

REFERENCES.....	54
APPENDICES.....	58
1. Appendix A. - Terms of Reference.....	58
2. Appendix B. - All Experimentation Code	69

List of Tables

Table 1. Demographic and experimental information about the subjects of the dataset	15
Table 2. Independent features of the dataset with demographic information	17
Table 3. Independent features of the dataset with information about the biomedical voice measurements.....	17
Table 4. Dependent features of the dataset.....	19
Table 5. Data Splitting	23
Table 6. Dropped Variables.....	25
Table 7. Main differences between the Boosted Trees model and the Convolutional Neural Network	26
Table 8. Boosted Trees performance.....	27
Table 9. Convolutional Neural Network performance	33
Table 10. Top variables according to SHAP Value for Boosted Trees and CNN	38
Table 11. Some PCA Loading correlations	41
Table 12. Summary of the Feature Selection Process	43
Table 13. Boosted Trees (Improved Version) performance	44

List of Figures

Figure 1. Relationships between the clinical, perceptual and acoustic characteristics of the voice changes in Parkinson's disease (Adapted from Ma, Lau and Thyagarajan, 2020).....	12
Figure 2. Sample of the dataset	20
Figure 3. Comparison of the Pearson and Spearman Correlation Coefficients of the features related to the target variable (motor_UPDRS)	24
Figure 4. Pair plot of some features related to the target variable (motor_UPDRS)	24
Figure 5. Instance of the training set transformed into an image.....	31
Figure 6. Instance of the validation set transformed into an image	31
Figure 7. Train Loss vs Validation Loss, CNN.....	34
Figure 8. Mean SHAP Value (Average impact on model output), Boosted Trees	36
Figure 9. SHAP Value (Impact on model output) – Global Impact, Boosted Trees ...	36
Figure 10. Mean SHAP Value (Feature Importance) – Global Impact, CNN	37
Figure 11. SHAP Value – Local Impact, CNN	38
Figure 12. Principal Component Analysis Loading Plot	40
Figure 13. Comparison of different models for choosing the right value of k for the k-Best Method	42
Figure 14. Mean SHAP Value (Average impact on model output) – Global Impact, Boosted Trees	45
Figure 15. SHAP Value (Impact on model output) – Global Impact, Boosted Trees .	46
Figure 16. Error message in the Google Collaboratory platform (September 25, 2024)	50
Figure 17. Error message in the Google Collaboratory platform (September 27, 2024)	50

Abstract

In the context of a regression task applied to biomedical speech measurements of subjects with early-stage Parkinson's Disease, the performance of the model and the knowledge about the most important features of the model are crucial for its widespread implementation among healthcare professionals in the detection and treatment of Parkinson's Disease. In this project, the machine learning model of Boosted Trees was applied to a numeric tabular data set, and the deep learning technique of convolutional neural networks was applied to images of 224x224 pixels, which were previously a numeric tabular dataset. The Boosted Trees model achieved a Mean Absolute Error of 1.21 for the training phase and 1.23 for the testing phase on the metric tabular dataset; meanwhile, the convolutional neural network achieved a Mean Absolute Error of 6.43 for the training phase and 6.52 for the testing phase on the image dataset. In addition, both models display 8 features in common in the top 10 features that have more impact on model prediction according to the Mean Shapley Value technique. The aim of providing healthcare professionals with an effective, robust, and easy-to-use technological tool, while at the same time figuring out the most relevant features in the context of biomedical speech measurements of subjects with early-stage Parkinson's Disease, was achieved through the machine learning model of Boosted Trees. Furthermore, the evidence suggests that changes in the hyperparameters and/or architecture of the convolutional neural network are probably required, or the transformation of the numeric tabular data into images may not have been optimal, negatively impacting the convolutional neural network's performance

Declaration

No part of this project has been submitted in support of an application for any other degree or qualification at this or any other institute of learning. Apart from those parts of the project containing citations to the work of others, this project is my own unaided work. This work has been carried out in accordance with the Manchester Metropolitan University research ethics procedures, and has received ethical approval number 68634.

Signed: Ramses Moreno De La Cruz

Date:04-10-2024

Acknowledgements

I really appreciate the support of my family, parents, and my advisor, Dr. Indranath Chatterjee, during the labyrinthine and knotty journey of my dissertation project. From the bottom of my heart, thank you very much to all of you!

Abbreviations

PD	Parkinson's Disease
CNN	Convolutional Neural Network
MAE	Mean Absolute Error
XAI	Explainable Artificial Intelligence
EDA	Exploratory Data Analysis
PCA	Principal Component Analysis
GPU	Graphical Processing Unit
UPDRS	Unified Parkinson's Disease Rating Scale
HNR	Harmonics-to-Noise Ratio
CSV	comma-separated values

CHAPTER 1. - INTRODUCTION

1.1 Project Overview

It is very common to see in supervised machine learning approaches that many features considered highly correlated according to a specific metric are discarded after applying Exploratory Data Analysis methods such as correlation matrix, Principal Component Analysis, scatter plots, pair plots, heatmaps, and feature importance analysis, or after applying preprocessing techniques such as feature engineering. This situation leaves the model with just a few variables for the processes of training, validation, testing, and deployment.

Considering that there are multidimensional relationships and complex interactions between features of the numeric tabular dataset derived from speech measurements of Parkinson's Disease patients that might be missed or overlooked by traditional machine learning models, as was briefly discussed in the previous paragraph, and having the potential to improve regression task performance and/or bring new insights about the relevant voice biomarkers for the detection and treatment of Parkinson's Disease, this project seeks to compare the regression performance and feature importance of some well-known machine learning models (i.e., a Boosted Trees model) and a deep learning model (i.e., a convolutional neural network), which will be applied using an unconventional approach of transforming the whole numeric tabular dataset input of speech measurements into images.

The application of a deep learning model using the numeric tabular dataset of voice measurements transformed into images aims to improve the performance, insights, and understanding of Parkinson's Disease by leveraging the intricate and complex relationships of the features, which are visually represented in a meaningful way in the images.

1.2 Aim and Objectives

1.2.1 Aim

The main aim of the project is to provide to healthcare professionals with an effective and easy-to-use technological tool based on machine learning, deep learning, and explainable artificial intelligence that sheds light on the way the models make decisions and describes the most important features related to speech measurements for understanding Parkinson's disease.

In essence, the provided limitations of many existing research studies based on machine learning and deep learning tools to improve the patient's quality of life encourage the development of this project. In other words, a new approach that looks not only for detection but also for tracking the progress of Parkinson's disease, even when symptoms are not obviously noticeable to neurologists or movement disorder specialists. Therefore, this project could represent a viable option or at least one of the first steps to help people in the prodromal phase or at risk of Parkinson's disease based on the presence of clinical features, genetic variants, or other biomarkers, but without an official clinical diagnosis of Parkinson's disease.

1.2.2 Objectives

- Conduct a literature survey and review in the following fields: Parkinson's Disease, speech measurements, vocal biomarkers, deep learning, machine learning, and explainable artificial intelligence.
- Exploratory data analysis and data processing of the dataset of people with Parkinson's Disease.
- Design, build, and run a machine learning (i.e., a Boosted Trees model) and deep learning model (i.e., a convolutional neural network) for a regression task using the dataset of patients with Parkinson's Disease. Then, the explainable artificial intelligence technique of Shapley Values will be applied to explain the regression task and identify the most relevant features for the regression algorithm.

1.2 Dissertation Project Structure

This research project was split in the following sections, accordingly to the specific stage of the project:

- Chapter 1 (Introduction): it offers a succinct overview of the research project along with the aims and objectives of it.
- Chapter 2 (Literature Review): it covers the approaches, applications, outputs, and findings of research projects published by different members of the scientific and academic community, which are deeply related to the aims and objectives of this dissertation.
- Chapter 3 (Data): it comprises all the information about the attributes and properties of the dataset used in this research project.
- Chapter 4 (Design, Implementation and Results): it encompasses the design, build, development, and application of the machine learning model (i.e., a Boosted Trees model) and deep learning model (i.e., a convolutional neural network) during the experimentation phase of this project with the dataset described in Chapter 3. In addition, it contains the breakdown, analysis, and assessment of the outcomes of the experimentation phase.
- Chapter 5 (Further Work): This chapter shows key ideas or lines of thought about potential research endeavours as a byproduct of this dissertation. Moreover, it describes some of the challenges faced during the development of this project.
- Chapter 6 (Conclusions): it presents the summary of the conclusions and evaluations of the research project.

CHAPTER 2. - LITERATURE REVIEW

2.1 Introduction

A literature review was conducted with the purpose of gaining deep insight and understanding of several academic and scientific subjects relevant to the dissertation topic. This survey includes a critical assessment of the ideas, methods, approaches, techniques, procedures, and results that have already been applied and obtained, respectively.

The key topics reviewed during the literature survey and discussed on detail in the next subsection were the following:

- Parkison's disease
- Speech measurements and vocal biomarkers
- Deep learning
- Machine learning
- Explainable artificial intelligence

2.2 Parkison's disease

2.2.1 Description, symptoms and risk factors

Parkinson's disease is a prevalent disorder of the central nervous system (Pawar, Bais, and Bhadule, 2024; Wang and Song, 2022), caused by the progressive loss of dopaminergic neurons in a region of the brain called the substantia nigra (NHS, 2022). This loss of nerve cells produces a reduction in dopamine in the brain, which is the chemical messenger responsible for transmitting signals between the substantia nigra and the corpus striatum to produce smooth, purposeful movement (National Institute of Neurological Disorders and Stroke, 2023). As a consequence of the progressive reduction of dopamine, the patient suffers from many of the non-motor and motor symptoms of this disease. It is important to remark that it is the second most common neurodegenerative disorder worldwide (Tracy *et al.*, 2020).

Some of the non-motor symptoms of Parkinson's disease are emotional changes (National Institute of Neurological Disorders and Stroke, 2023), mild memory and

thinking problems, anxiety, dementia, depression, hallucinations, delusions (Parkinson's UK, 2020), loss of sense of smell, and problems sleeping (NHS, 2022). The main motor symptoms of Parkinson's disease are the following (National Institute of Neurological Disorders and Stroke, 2023):

- Tremor: It begins in a hand, a foot, or the jaw. This symptom is evident when the hand is at rest or when a person is under stress, and in general, it disappears during sleep or improves with an intended movement.
- Rigidity: This symptom is characterized by muscle stiffness. The person's muscles remain constantly tense and contracted. Consequently, there is a certain resistance to movement.
- Bradykinesia: This symptom is described as the slowing down of spontaneous and automatic movement. Therefore, tasks that were simple and easy become difficult and take longer. A decrease in facial expressions is also exhibited because of this symptom.
- Postural instability: It is reflected by impaired balance and changes in posture.

It is extremely important to remark that Parkinson's disease symptoms typically begin on one side of the body. Nonetheless, it finally affects both sides of the body, although symptoms are often less severe on one side than on the other (National Institute of Neurological Disorders and Stroke, 2023).

Some risk factors for Parkinson's disease include the following (National Institute of Neurological Disorders and Stroke, 2023):

- Age: The average age of beginning of this disease is about 70 years. However, there is evidence of people who developed the disease before the age of 50.
- Biological sex: Parkinson's disease affects more men than women.
- Heredity: An estimated 15 to 25 % of people with Parkinson's disease have a known relative with the disease.
- Exposure to pesticides.

2.2.2 Diagnosing Parkinson's disease

Currently, there is not a specific single approved test that diagnoses Parkinson's disease; therefore, the diagnosis is based on the combinatorial use of some of the following well-known medical tests and procedures with the proper guidance of a neurologist or movement disorders specialist (National Institute of Neurological Disorders and Stroke, 2023):

- Medical history and a neurological examination, such as postural instability, tremor, balance abnormalities, rigidity and facial expressions.
- Blood and laboratory tests.

- Brain scans such as computed tomography and magnetic resonance imaging.
- Genetic tests.
- Physiological observations correlated with the auditory phenomena, such as Laryngoscopy, Photoglottography, Laryngeal EMG, Laryngeal CT, Respiratory assessment (Ma, Lau and Thyagarajan, 2020).
- Qualitative voice examinations, also called Perceptual Analysis based on the Unified Parkinson's Disease Rating Scale and experienced listeners (Suppa *et al.*, 2022).

In the context of an elusive single biomarker to clinically diagnose Parkinson's disease, it is worth mentioning that in the last few years, there have been several research and experimental projects that have the potential to dramatically change the management (Tracy *et al.*, 2020) and cost associated with the detection and diagnosis of Parkinson's disease, but they are still not considered a standard medical procedure or tool. Those methods are the following:

- Voice analysis based on Machine Learning and Deep Learning: it is an objective (Ma, Lau, and Thyagarajan, 2020) and non-invasive procedure that extracts and analyses acoustic properties of the voice to detect or diagnose, with high speed, accuracy, and precision, the status of Parkinson's disease in a patient (Hecker *et al.*, 2022; Tracy *et al.*, 2020). This technique, which has the capability to reach hundreds of millions of people at low or almost zero cost, is the experimental procedure that provides the dataset that will be studied in this dissertation project
- Skin biopsy: it is a minimally invasive skin biopsy procedure that uses three 3-millimeter skin punch biopsies taken from the neck, the knee, and the ankle of the patient. Then the procedure looks for the presence of the abnormal protein called phosphorylated α -synuclein in the nerve fibers in the skin. If the protein is present, then someone is considered a carrier of neurodegenerative disorders or synucleinopathies, of which Parkinson's disease is one (Gibbons *et al.*, 2024).

2.2.3 Treating Parkinson's disease

Currently, there is no cure for Parkinson's disease, but medications, surgery, deep brain stimulation, and complementary therapies such as massages, yoga, hypnosis, acupuncture, and emotional support can provide some control over the motor and/or non-motor symptoms. However, it is relevant to remark that the symptoms may gradually return over time as Parkinson's disease progresses and worsens, and the drugs or other medical procedures become less effective (National Institute of Neurological Disorders and Stroke, 2023).

2.3 Voice disorders

Voice disorders are considered one of the earliest signs of motor symptoms in Parkinson's disease, specifically an indicator of the bradykinesia symptom (Ma, Lau and Thyagarajan, 2020; Wang and Song, 2022; Tracy *et al.*, 2020). Some studies suggest that changes in the voice can manifest 5 to 7 years before the official diagnosis of Parkinson's disease (Ma, Lau and Thyagarajan, 2020). Specifically, variability in fundamental frequency has been detected as early as 5 years before the onset of clinically diagnosable symptoms (Tracy *et al.*, 2020). These changes in the voice appear to stem from two main pathological processes in the larynx: asymmetric rigidity of the intrinsic laryngeal muscles and incomplete glottic closure due to vocal cord hypokinesis (Ma, Lau and Thyagarajan, 2020).

It is estimated that 70–89% of Parkinson's disease patients will experience vocal impairment as the disease progresses, and greater than 30% of these patients find it to be the most bothersome of all their symptoms (Wang and Song, 2022). Patients often have difficulty initiating speech, speak in short rushes of words, have inappropriate silences, have a variable speaking rate, and can have a vocal tremor (Wang and Song, 2022). Voice changes or hypokinetic dysarthria occur in key components of speech production such as phonation, articulation, and prosody in the following way (Ma, Lau, and Thyagarajan, 2020):

- Changes in phonation: there is a reduction in the voice volume (hypophonia) and altered voice quality (dysphonia).
- Changes in articulation: there is a reduction in the range of articulatory movements (hypokinetic articulation).
- Changes in prosody: it is also called dysprosody of speech and it is manifested by flattened pitch inflection (monopitch) and loss of stress (monoloudness).

The level of complexity of the phonation, prosody and articulation process is enormous considering the number and type of body structures involved. To be precise, they work in the following way:

- Phonation: This process involves producing sound waves through the vibration or cycles of opening and closing of the vocal cords within the larynx due to the movement of air from the lungs. It is important to remark that faster vibrations of the vocal cords produce a higher-pitched sound, and conversely, slower vibrations of the vocal cords produce a lower pitch. In addition, the volume of the sound wave is influenced by the force of the air passing through the vocal cords.

- Articulation: This process involves the transformation of the sound waves produced by the vocal cords during the phonation process into specific speech sounds, such as consonants and vowels. This refinement process encompasses articulators such as the tongue, teeth, lips, hard palate, and soft palate (The Voice Foundation, 2019).
- Prosody: It involves the study of rhythm, stress, and intonation in speech. It focuses on syllables and larger speech segments rather than individual phonetic segments like vowels and consonants (Tatham and Morton, 2006).

2.4 Perceptual Analysis and Acoustic Analysis

According to the perceptual analysis practiced by experienced listeners and the acoustic analysis generated by machine learning and deep learning models, these are some of the most important voice features impacted by Parkinson's disease (Ma, Lau, and Thyagarajan, 2020):

- Changes in quality:
 - Perceptual analysis: it shows a rough, weak and breathy voice quality.
 - Acoustic analysis:
 - A reduced harmonics-to-noise ratio means a rough and hoarse voice. The harmonics-to-noise ratio measures the proportion of harmonic sound derived from the vibration of the vocal cords against the noise generated from the glottis.
 - Unstable vocal cord vibration is associated with a rough vocal quality, where the presence of roughness signifies turbulent airflow, which can be a sign of aperiodic vibration of the vocal folds (Wang and Song, 2022). In addition, an increase in jitter has been linked with Parkinson's disease. Jitter is a measure of the cycle-to-cycle variability in the fundamental frequency during steady phonation.
 - Breathiness is caused by glottic air leakage, which adds noise because of turbulent airflow in the larynx. In addition, an increase in shimmer has been linked with Parkinson's disease. Shimmer is a measure of the cycle-to-cycle variation in amplitude during steady phonation.
- Changes in volume:
 - Perceptual analysis: Impairments in volume control producing hypophonia and monoloudness.
 - Acoustic analysis: It is reflected as reduce range and variation of intensity of the volume.

- Changes in pitch:
 - o Perceptual analysis: It reports higher pitch in the early stages of Parkinson's disease. In addition, it reports a monopitch character with reduced intonation of voice, which corresponds to a decrease in the variability of fundamental frequency in acoustic analysis.
 - o Acoustic analysis: Higher pitch is correlated with higher mean fundamental frequency in many individuals with Parkinson's disease. Fundamental frequency is clinically perceived as pitch.
- Summary (see Figure 1)

Speech Domain	Clinical and Perceptual Features	Acoustic Features
Phonation	Dysphonia	Decreased harmonics-to-noise ratio
	Rough/asthenic voice	Increased jitter
	Higher pitch	Increased mean F_0
	Hypophonia	Reduced range of intensity
	Breathy voice	Increased shimmer
Prosody	Monoloudness	Decreased variation of intensity
	Monopitch	Decreased variability of F_0

Figure 1. Relationships between the clinical, perceptual and acoustic characteristics of the voice changes in Parkinson's disease (Adapted from Ma, Lau and Thyagarajan, 2020)

2.5 Survey of Machine and Deep Learning in Acoustic Analysis

Recent research studies employing machine learning algorithms in extracted voice features to classify Parkinson's disease from non-Parkinson's reported high classification accuracy around 98%-99%; however, they have the following limitations (Tracy *et al.*, 2020):

- Data size: It undermines the generalizability of the results to a larger and more diverse population.
- Identity confounding: This occurs when multiple voice samples are collected from each individual, and these samples appear in both the training and testing data. This leads to over-optimistic model performance because the model has learned to identify characteristics of specific individuals and is using that information to predict the label in the test set. In essence, this is a type of data leakage where information is unintentionally shared between the training and test sets, leading to inflated performance metrics.

Using non-invasive, easily obtained, low-cost, patient-generated voice data, this dissertation project outlined voice analysis methodologies related to a machine learning model (i.e., Boosted Trees), a deep learning model (i.e., a convolutional neural network), and the transformation of the numeric tabular dataset into images along with the Explainable Artificial Intelligence methodology of Shapley Value.

This project would use biomedical speech measurements of subjects with early-stage Parkinson's disease to gauge the application of the methodologies previously listed. One of the main contributions of this dissertation project is its relevance for the identification of the most important features for the diagnosis of Parkinson's disease. In addition, the approach of this project extends the general understanding of regression using a convolutional neural network on images that were previously a numeric tabular dataset, which has the potential to greatly improve quality of life and clinical outcomes considering the multiple and complex relationships that could be revealed. We aren't aware of any other studies that have attempted to use a convolutional neural network on images that were previously a numeric tabular dataset of biomedical speech measurements of subjects with early-stage Parkinson's disease for a regression task.

CHAPTER 3. - DATA

3.1 Introduction

After describing in Chapter 2 relevant research endeavours associated with Parkinson's disease and biomedical voice measurements, Chapter 3 elucidates significant components of the dataset that was used in this dissertation project. The aspects of the dataset of biomedical speech measurements of subjects with early-stage Parkinson's disease that are going to be presented in this chapter are related to the source of the data, data format, data type, demographic and clinical information of the subjects or participants that generated the raw data, dependent and independent features, missing data, balanced dataset, splitting the dataset, and a sample of the dataset.

3.2 Key insights about the dataset

- Summary of the dataset:
 - o The dataset contains biomedical voice measurements from subjects with early-stage Parkinson's disease recruited to a six-month trial of a telemonitoring device for remote symptom progression of Parkinson's disease (Kancharla Naveen Kumar, 2023; *UCI Machine Learning Repository*, no date).
- Format of the dataset:
 - o The dataset is in comma-separate value format.
- Source of the dataset:
 - o Kaggle - Parkinson's Disease Detection Dataset
 - <https://www.kaggle.com/datasets/naveenkumar20bps1137/parkinsons-disease-detection/data?select=parkinsons.data>
 - License: Creative Commons Attribution 4.0 International (CC BY 4.0).
 - o UC Irvine Machine Learning Repository
 - <https://archive.ics.uci.edu/dataset/174/parkinsons>
 - License: Creative Commons Attribution 4.0 International (CC BY 4.0).

- Background of the dataset (Tsanas *et al.*, 2010):
 - o The raw data or audio recordings that were used to build this dataset were collected using a telemonitoring system, which includes a high-quality microphone headset for recording, making it easier to remotely measure several motor impairment symptoms of Parkinson's disease from subjects' homes. Then, the collected raw data from the subjects' homes was transmitted online to the clinic to be processed.
 - o The audio recordings (i.e., the raw data) of the subjects were of two types:
 - Sustained phonation of the vowel "ahhh..." to avoid the confounding effects of running speech, consequently simplifying the signal analysis.
 - Running speech in which the subject describes a static photograph.
 - o The audio recordings were made using a high-quality microphone headset placed 5 cm from the patient's lips.
- Subjects of the dataset:
 - o The subjects of this dataset were diagnosed with Parkinson's disease if they had at least two of the following symptoms: rest tremor, bradykinesia or rigidity. The diagnosis was within the previous five years at trial onset (Tsanas *et al.*, 2010).
 - o Demographic and experimental information about the subjects of the dataset is described in the Table 1.

Table 1. Demographic and experimental information about the subjects of the dataset

	Subjects (unit)	Feature "age" of the subjects (years)			Recordings of the subjects (unit)			
		Quantity	Maximum Value	Mean Value	Minimum Value	Quantity	Maximum Value	Mean Value
All	42	85	64.80	36	5875	168	140	101
Male	28	78	65.05	49	4008	168	143	107
Female	14	85	64.27	36	1867	165	133	101

- Data type:
 - o The data type in this dataset is only metric; in other words, it encompasses integers and decimals.
- Missing data:
 - o There is no missing data or null values in the dataset used in this project.
- Data Augmentation:

- This technique was not implemented in this project.
- Balanced dataset:
 - The dataset used in this project is not considered unbalanced because the 2:1 ratio displayed in this dataset (i.e., two males for every female) reflects the distribution in the general population for Parkinson's disease. According to Crispino et al. (2020), the ratio between men and women is about 2:1 at the time of diagnosis of Parkinson's disease, probably due to the neuroprotective role of estrogen against degenerative damage in females. Consequently, females show fewer symptoms in the preclinical stage of Parkinson's disease compared to men. However, once Parkinson's disease has reached a more advanced and established clinical stage, no more gender differences can be observed. In addition, the Parkinson's Foundation (2024) indicates that men are 1.5 times more likely to have Parkinson's disease than women. Furthermore, according to Cerri, Mus, and Blandini (2019), increasing evidence points to biological sex as an important factor in the development of Parkinson's disease. The risk of developing Parkinson's disease is twice as high in men as in women, but women have a higher mortality rate and faster progression of the disease.
- Data distribution used in this project:
 - 60 % of the dataset for the training set
 - 20 % of the dataset for the validation set
 - 20 % of the dataset for the test set

3.3 Description of the Dependent and Independent variables

- Rows and columns of the dataset:
 - The rows of the dataset carry the voice recordings, and the columns of the dataset have the independent variables related to the biomedical voice measurement and demographic information about the subjects of the trial.
 - Total number of rows of the dataset (i.e., total number of individual recordings of the dataset): 5875 rows
 - Columns of the dataset
 - Independent features: 20 columns, where 17 correspond to biomedical voice measurement and 3 correspond to demographic information.
 - Dependent feature: 2 columns

- Independent features of the dataset with demographic information (see Table 2)

Table 2. Independent features of the dataset with demographic information

Nº	Feature	Description
1	Subject	integer that uniquely identifies each subject.
2	Age	age of the subject, in years.
3	Sex	gender of the subject, where the number 0 represents a male and the number 1 represents a female.

- Independent features of the dataset with information about the biomedical voice measurements (see Table 3)

Table 3. Independent features of the dataset with information about the biomedical voice measurements

Nº	Feature	Description
4	test_time	time since recruitment into the trial, in days (Kanchala Naveen Kumar, 2023; UCI Machine Learning Repository, no date).
5	Jitter (%)	<p>Jitter is a parameter of voice frequency perturbation (Teixeira and Gonçalves, 2014).</p> <p>Jitter (%) is a parameter or measure of Jitter.</p> <p>Jitter (%): it represents the average absolute difference between consecutive periods of vocal fold vibration, divided by the average period, expressed as a percentage. It quantifies cycle-to-cycle variations in the fundamental frequency (Kiran, Mounika, and Srinivas, 2015). In other words, it measures in percentage how much the pitch changes.</p> <p>The fundamental frequency corresponds to the lowest frequency produced by the vibration of the vocal cords, and this vibration of the vocal cords creates the basic pitch of the voice (Bowen <i>et al.</i>, 2014).</p>
6	Jitter (Abs)	<p>Jitter (Abs) is a parameter or measure of Jitter.</p> <p>It is the absolute jitter. It represents the average absolute difference between pitches. It means that it measures the deviation of fundamental frequency from one pitch period cycle to the other (Kiran, Mounika and Srinivas, 2015).</p>
7	Jitter: RAP	<p>Jitter: RAP is a parameter or measure of Jitter.</p> <p>The Relative Average Perturbation (RAP) is the average absolute difference between a period and the average of it and its two neighbours, divided by the average period (Teixeira and Gonçalves, 2014). Strictly speaking, it measures how the pitch changes or pitch disturbances over three short voice samples or periods (Kiran, Mounika and Srinivas, 2015).</p>
8	Jitter: PPQ5	Jitter: PPQ5 is a parameter or measure of Jitter.

		The Five-Point Period Perturbation Quotient (PPQ5) is the average absolute difference between a period and the average of it and its four closest neighbours, divided by the average period (Teixeira and Gonçalves, 2014). It measures how the pitch changes or pitch disturbances over five short voice samples or periods (Kiran, Mounika, and Srinivas, 2015).
9	Jitter: DDP	<p>Jitter: DDP is a parameter or measure of Jitter.</p> <p>The Difference of Differences between Periods (DDP) is the average absolute difference between consecutive differences between consecutive periods, divided by the average period (Little <i>et al.</i>, 2009). It is another way to measure pitch changes over three samples.</p>
10	Shimmer	<p>Shimmer is a parameter of voice amplitude perturbation (Tsanas <i>et al.</i>, 2010).</p> <p>It measures the cycle-to-cycle variations or consecutive period variations of amplitude in the voice (Teixeira and Gonçalves, 2014).</p> <p>The shimmer changes, or more instability in vocal amplitude, to be specific, an increase in shimmer values is usually observed in subjects with Parkinson's disease (Kiran, Mounika, and Srinivas, 2015; Teixeira and Gonçalves, 2014).</p>
11	Shimmer(dB)	<p>Shimmer(dB) is a parameter or measure of Shimmer.</p> <p>Shimmer expressed in decibels.</p>
12	Shimmer: APQ3	<p>Shimmer: APQ3 is a parameter or measure of Shimmer.</p> <p>The Amplitude Perturbation Quotient 3 (APQ3) is the measure of the average absolute difference between the amplitude of a period and the average of the amplitudes of its two closest neighbours, divided by the average amplitude.</p>
13	Shimmer: APQ5	<p>Shimmer: APQ5 is a parameter or measure of Shimmer.</p> <p>The Amplitude Perturbation Quotient 5 (APQ5) is the measure of the average absolute difference between the amplitude of a period and the average of the amplitudes of its four closest neighbours, divided by the average amplitude.</p>
14	Shimmer: APQ11	<p>Shimmer: APQ11 is a parameter or measure of Shimmer.</p> <p>The Amplitude Perturbation Quotient 11 (APQ11) is the measure of the average absolute difference between the amplitude of a period and the average of the amplitudes of its ten closest neighbours, divided by the average amplitude.</p>
15	Shimmer: DDA	Shimmer: DDA is a parameter or measure of Shimmer.

		The Directional Difference Average (DDA) measures the average absolute difference between consecutive differences in the amplitudes of consecutive periods (Tsanas <i>et al.</i> , 2010).
16	NHR	The Noise-to-Harmonics Ratio (NHR) measures the amount of noise (i.e., hoarseness) in the voice signal compared to the periodic components (Williamson, 2014).
17	HNR	The Harmonics-to-Noise Ratio (HNR): It is the inverse of NHR. It measures the ratio between periodic and non-periodic components of a speech sound (Fernandes <i>et al.</i> , 2018). It could indicate the presence of nonharmonic phenomena, which are tied to the perception of hoarseness or breathiness (Ikuma <i>et al.</i> , 2022).
18	RPDE	The Recurrence Period Density Entropy (RPDE) quantifies the ability of the vocal folds to sustain simple vibration, measuring the deviations from exact periodicity (Tsanas <i>et al.</i> , 2010). Strictly speaking, it measures the irregularities in the voice.
19	DFA	The Detrended Fluctuation Analysis (DFA) quantifies the stochastic self-similarity of the noise in the voice caused by turbulent airflow in the vocal tract (Tsanas <i>et al.</i> , 2010). This means that changes in DFA can indicate alterations in voice turbulence.
20	PPE	The Pitch Period Entropy (PPE) measures the impaired control of stable pitch during sustained phonation (Tsanas <i>et al.</i> , 2010). Which is to say, that PPE measures how much the pitch changes.

- Dependent features of the dataset are described in the Table 4

Table 4. Dependent features of the dataset

Nº	Feature	Description
21	motor_UPDRS	The Motor UPDRS (Unified Parkinson's Disease Rating Scale) assesses motor and aspects of Parkinson's disease in the patients; therefore, it provides relevant information about the disease progression and their actual damages.
22	total_UPDRS	The Total UPDRS (Unified Parkinson's Disease Rating Scale) assesses both motor and non-motor aspects of Parkinson's disease in the patients (Bárcenas, Fuentes-García and Naranjo, 2022); therefore, it provides relevant information about the disease progression and their actual damages.

- A sample of the dataset (see Figure 2)

Head of the dataset:								
subject#	age	sex	test_time	motor_UPDRS	total_UPDRS	Jitter(%)	\	
0	1	72	0	5.6431	28.199	34.398	0.00662	
Jitter(Abs)	Jitter:RAP	Jitter:PPQ5	...	Shimmer(dB)	Shimmer:APQ3	\		
0	0.000034	0.00401	0.00317	...	0.23	0.01438		
Shimmer:APQ5	Shimmer:APQ11	Shimmer:DDA		NHR	HNR	RPDE	DFA	\
0	0.01309	0.01662	0.04314	0.01429	21.64	0.41888	0.54842	
PPE								
0	0.16006							

Figure 2. Sample of the dataset

CHAPTER 4. - DESIGN, IMPLEMENTATION AND RESULTS

4.1 Introduction

After detailing in Chapter 3 pertinent information about the characteristics and conditions of the dataset, Chapter 4 explains in detail the most crucial aspects of the logical design and implementation process of the regression models in which the dataset outlined in Chapter 3 will be applied.

Chapter 4 consists of the design, building, and execution of two regression models; one of them is Boosted Trees and the other is a Convolutional Neural Network, with the purpose of comparing the regression performance with metrics such as Mean Absolute Error, Mean Square Error, and Root Mean Square Error, and gaining deeper and better insights into the most relevant features for healthcare professionals from a dataset of biomedical speech measurements of subjects with early-stage Parkinson's disease using the Explainable Artificial Intelligence technique called Shapley Values.

It is important to note that the Boosted Trees model will use the dataset of biomedical speech measurements of subjects with early-stage Parkinson's disease in tabular format, while the Convolutional Neural Network will use the same tabular dataset, but transformed into images.

4.2 Set up the Environment / Loading the dataset

4.2.1 Set up the Environment

Setting up the environment and importing the respective libraries, functions, modules, and frameworks related to the Python programming language.

Firstly, it is important to remark that all code experiments were built and executed using the cloud-based platform Google Colaboratory, also called Google Colab (version 2024), running Python (version 3.10.12). The decision to build the code of the dissertation project using Google Colaboratory was based on the fact that this platform permits the

user to write and execute Python code directly in the web browser. In addition, Google Colaboratory allows seamless execution of computationally intensive regression tasks, including the training and testing of machine learning and deep learning models such as Boosted Trees and Convolutional Neural Networks, respectively, harnessing its pre-installed libraries such as PyTorch and scikit-learn.

Additional relevant information about the set-up process for developing the dissertation project are as follows:

- Computational Resources:
 - o Laptop
 - 64-bit operating system, x64-based processor
 - Installed RAM: 8.00 GB
 - Processor: 11th Gen Intel(R) Core (TM) i5-1135G7 @ 2.40GHz
 - Sockets: 1
 - Cores per socket (physical cores): 4
 - Logical cores: 8
- Data Storage and Access: The code notebook and the dataset loaded in it was stored on Google Drive and accessed using Google Colaboratory.

4.2.2 Loading and having a glance of the dataset.

- Data used: whole dataset (all features).

4.3 Splitting the whole dataset

- Data used: whole dataset (all features).

The splitting of the dataset in a stratified way followed a determined proportion for the training set, validation set, and test set (see Table 5). In other words, a specific amount of the biomedical voice measurements of every subject was distributed among the training, validation, and test sets using the unique values of the column “subject” as a parameter for the splitting of the whole dataset, which also serves as a way to anonymously identify every participant in the dataset.

Table 5. Data Splitting

	Percentage of instances from the whole dataset	Number of instances from the whole dataset
Data Subset		
Training set	~ 60%	3584
Validation set	~ 20%	1145
Test set	~ 20%	1146
Total	100%	5875

4.4 Exploratory Data Analysis

- Data used: stratified training set (all features).

The Exploratory Data Analysis of the training set was done through the computation and display of the respective univariate analysis, such as bar plots and histograms for all variables. Furthermore, the appropriate bivariate analysis, including pair plots, joint plots, scatterplots, heatmaps, Pearson Correlation Coefficient, and Spearman Correlation Coefficient, was applied to understand the relationship between the independent features and the target variable (i.e., the variable “motor_UPDRS”).

From the univariate analysis, the main insight is that more than fifty percent (50%) of the independent variables have a histogram with a right-skewed data distribution, and only a few independent variables have a histogram with a normal-shaped data distribution.

Some of the key takeaways from the bivariate analysis between the independent features and the target variable (i.e., the variable “motor_UPDRS”) based on the Pearson Correlation Coefficient and Spearman Correlation Coefficient (see Figure 3) point out the following:

- “age” and “PPE” (Pitch Period Entropy) are some of the independent features with the highest positive values of correlation with the target variable (i.e., the variable “motor_UPDRS”).
- “HNR” (Harmonics-to-Noise Ratio) and “DFA” (Detrended Fluctuation Analysis) are the independent features with the highest negative values of correlation with the target variable (i.e., the variable “motor_UPDRS”).
- Despite the fact that the variables listed previously (“age”, “PPE”, “HNR”, “DFA”) have the highest values of correlation with the target variable (i.e., the variable “motor_UPDRS”) in both Pearson and Spearman Correlation Coefficients, none

of them are above 0.31. Consequently, the correlation between the independent variables and the target variable is small enough to conclude that the linear relationship between those is weak. Furthermore, the graphs such as pair plots and scatterplots between the independent variables and the target variable (see Figure 4) provide additional evidence that reinforce the statement that the linear relationship is weak.

Comparison of Pearson and Spearman Correlation Coefficient			
	Metric Variable	Pearson Correlation	Spearman Correlation
0	motor_UPDRS	1.000000	1.000000
1	total_UPDRS	0.947476	0.957649
2	age	0.282680	0.312043
3	subject	0.209668	0.197384
4	PPE	0.151383	0.149625
5	Shimmer:APQ11	0.116734	0.142993
6	RPDE	0.103239	0.122358
7	Shimmer(dB)	0.091120	NHR
8	Shimmer	0.084116	Shimmer
9	Jitter(%)	0.074269	Jitter(%)
10	Shimmer:APQ5	0.072014	Jitter:PPQ5
11	test_time	0.069940	Shimmer:APQ5
12	Jitter:PPQ5	0.069145	Shimmer:APQ3
13	Shimmer:APQ3	0.066514	Shimmer:DDA
14	Shimmer:DDA	0.066514	Jitter:DDP
15	NHR	0.066099	Jitter:RAP
16	Jitter:DDP	0.063351	RPDE
17	Jitter:RAP	0.063341	test_time
18	Jitter(Abs)	0.039854	Jitter(Abs)
19	sex	-0.050313	sex
20	HNR	-0.133199	HNR
21	DFA	-0.133771	DFA

Figure 3. Comparison of the Pearson and Spearman Correlation Coefficients of the features related to the target variable (motor_UPDRS)

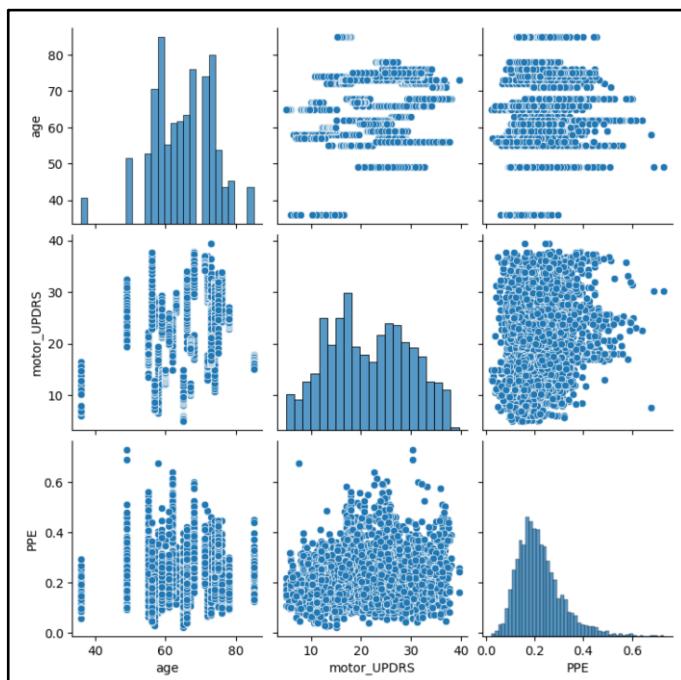


Figure 4. Pair plot of some features related to the target variable (motor_UPDRS)

4.5 Feature Engineering

In this phase, the pipelines to make the transformation or imputation process of variables more straightforward were built. It is important to remark that this phase was extremely important for developing the Boosted Tree model. Moreover, some of the variables of the dataset that were not used during the several stages of model building and development were dropped, as listed and explained in Table 6.

Table 6. Dropped Variables

Variable name	Explanation for dropping	
subject	Type of variable:	- Independent variable. - Metric variable.
	Explanation for dropping:	Considering the fact that the variable “subject” only shows an integer to uniquely and anonymously identify each subject of the database, the usefulness of the variable “subject” is null in the context of developing a robust linear regression model for the prediction of the variable “motor_UPDRS.”
total_UPDRS	Type of variable:	- Dependent variable. - Metric variable.
	Explanation for dropping:	Considering that the variable “total_UPDRS” assesses both motor and non-motor symptoms of Parkinson’s disease, and the variable “motor_UPDRS” assesses exclusively motor symptoms of Parkinson’s, which include voice changes, the variable “total_UPDRS” is not directly related to motor-specific assessments as the variable “motor_UPDRS”; therefore, the variable “total_UPDRS” was discarded as the target variable at the moment of building the regression models.

4.6 Contrasts of the Boosted Trees model and the Convolutional Neural Network

The evaluation of the performance metrics and feature importance of the Boosted Trees and the Convolutional Neural Network using the dataset of biomedical speech measurements of subjects with early-stage Parkinson’s disease was done, taking into account the limitations and considerations briefly described in Table 7 and detailed in Sections 4.7, 4.8, and 4.9:

Table 7. Main differences between the Boosted Trees model and the Convolutional Neural Network

	Boosted Trees	Convolutional Neural Network
Input Format	Tabular data (numeric features)	Images generated from the tabular data (numeric features)
Model Architecture	It performs well on structured data (tabular data) and can capture feature interactions and non-linear relationships.	It performs well on spatial data and can learn feature representations from the image.

4.7 Boosted Tree Model

4.7.1 Boosted Tree Model - Description

The boosted trees model for a regression task is an ensemble machine learning model that combines multiple weak decision trees, where each new tree is trained to predict the differences between predicted and actual values of the current ensemble. Consequently, this ensemble model builds a powerful model after every iteration, where the final prediction is the weighted sum of the predictions from all the trees. Looking for the right balance between performance and interpretability, the machine learning model selected for developing this project is Boosted Trees. In detail, the reasons for choosing this model are the following:

- It can capture complex non-linear relationships compared to linear models, which is extremely useful considering the weak levels of linear relationship between the independent variables and the target variable according to the information provided by “Section 4.4 Exploratory Data Analysis.”
- It does not require specifying polynomial terms like Polynomial Models with the goal of capturing intricate patterns.
- It has great predictive performance and robustness to overfitting compared to Random Forest models.
- This kind of model is formidable for handling tabular or structured data.

4.7.2 Boosted Tree Model - Training Phase, Validation Phase and Test Phase

- Data used: stratified training set, stratified validation set, stratified test set (all features)

The training, validation and test phase using the Boosted Trees model considered as base the pipelines developed in the “Section 4.5 Feature Engineering”. The summary of

the key outputs of the Boosted Trees model in the training set, validation set, and test set are shown in the Table 8.

Table 8. Boosted Trees performance

Phase	Model	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Mean Square Error (MSE)	Coefficient of Determination (R^2)	Features
Training (Dataset: stratified training dataset)	Boosted Trees [max_depth: 5]	1.21	1.66	2.77	0.96	"age", "sex", "test_time", "Jitter (%)", "Jitter (Abs)", "Jitter: RAP", "Jitter: PPQ5", "Jitter: DDP", "Shimmer", "Shimmer(dB)", "Shimmer: APQ3", "Shimmer: APQ5", "Shimmer: APQ11", "Shimmer: DDA", "NHR", "HNR", "RPDE", "DFA", 'PPE'
Validation Grid Search Cross validation (Dataset: stratified validation dataset)	Best Model ↓ Boosted Trees [max_depth: 10, n_estimators: 110]	7.46	-	-	-	
Test (Dataset: stratified test dataset)	Best Model ↓ Boosted Trees [max_depth: 10, n_estimators: 110]	1.23	2.59	6.70	0.90	

4.7.3 Boosted Tree Model – Analysis and Evaluation of the results

The difference between the Mean Absolute Error of the Boosted Trees model using the stratified training set (MAE = 1.21) and the Mean Absolute Error of the Boosted Trees fine-tuned with the best parameters in the stratified test set (MAE = 1.23) is significantly small (see Table 8). Therefore, it is irrefutable the power and advantages of the Boosted

Trees model in the context of prediction performance. This is probably due to the fact that the boosted tree model is well suited to address scenarios in which there are non-linear and/or complex relationships between the variables of the dataset.

Furthermore, this Boosted Trees model implemented could be considered a complex model because of the number of variables it has (more than fifteen), which tends to increase the Coefficient of Determination, even if those variables are not truly significant. Consequently, a great number of variables makes the Boosted Trees model prone to overfitting the dataset, as it is reflected in the Coefficient of Determination (see Table 8) during the training phase ($R^2 = 0.96$) and test phase ($R^2 = 0.90$).

4.8 Convolutional Neural Network

4.8.1 Convolutional Neural Network – Description

A convolutional neural network is a deep learning model built to process spatial data (i.e., images) by learning hierarchical feature representations through a series of convolutional layers (Goodfellow, Bengio, and Courville, 2016). In this project, the convolutional neural network is going to be used for a regression task to predict the value of the target variable; in other words, the variable “motor_UPDRS.”

It is crucial to note that the original format of the dataset was structured or tabular; therefore, before using that numeric dataset in the convolutional neural network, the whole tabular dataset was transformed into images using the algorithm called “Image Generator for Tabular Data.” Further explanations about this algorithm are described in “Section 4.8.2 Image Generator for Tabular Data – Description.”

Looking for the right balance between performance and interpretability, the deep learning model selected for developing this project is a convolutional neural network. In detail, the reasons for choosing this model are the following:

- Probably, it can capture better than the Boosted Trees model non-linear patterns or complex interactions, and spatial relationships among the features placed in the images as pixels, which is extremely useful considering the weak levels of linear relationship between the independent variables and the target variable according to the information provided by the “Section 4.4 Exploratory Data Analysis”. It is essential to note that a convolutional neural network learns

hierarchical feature representations through a conjunction of convolutional layers, pooling, and fully connected layers.

- Considering the architecture of this kind of model, it can handle high-dimensional data (i.e., a great number of features) capturing complex feature interactions in an efficient way.

In the context of this project, it is pertinent to remark that the convolutional neural network designed, built, and developed in this project was a vanilla convolutional neural network because of the reasons explained as follows:

- The architecture was tailored for the specific regression task.
- It was faster and computationally efficient for training and inference compared to the vast majority of the current pre-trained convolutional neural networks.

It reduced the engineering complexities related with the fine-tuning strategies.

4.8.2 Image Generator for Tabular Data – Description

- Data used: stratified training set, stratified validation set, stratified test set (all features)

The use of a convolutional neural network that has as input a dataset of images that were previously a tabular dataset was done in this project taking into account the following considerations:

- There is a certain level of engineering complexity at the moment to do the transformation of the dataset from the tabular format to images due to feature arrangement, pixel scaling, and normalization (Zhu *et al.*, 2021). Consequently, the transformation process of the dataset has a great influence on the performance of the convolutional neural network.
- Maybe the features of the tabular dataset do not have an evident spatial relationship that can be encoded properly into an image and consequently harnessed by the convolutional neural network. This means that probably the performance of the convolutional neural network is not going to be as good or even better than the performance of the Boosted Trees.
- A lot of computational resources and time are implemented to do the transformation from the tabular dataset to images.

The algorithm “Image Generator for Tabular Data” was used to transform the tabular data into images, and then those images were passed as input to the convolutional neural networks. Some key details of this algorithm are described in the following lines:

- The reconfiguration of the tabular dataset into images is based on the arrangement of the features of the tabular dataset into a matrix format where spatial relationships can be introduced, with each pixel position representing a

feature (i.e., this algorithm provides a compact image representation) and the pixel intensity displaying the value of the feature (Zhu *et al.*, 2021).

- The process of assigning a feature to a pixel position is the first main step of this algorithm
- In the matrix where the features are arranged, which is called the similarity matrix, the features that are more correlated or have similar relationships are placed closer together. As a result, the convolutional neural networks can exploit local spatial dependencies between features (Zhu *et al.*, 2021).
 - The second main step of this algorithm consists of quantifying the feature similarity (i.e., the pairwise distances between features or the distance between features in a given space). This distance could be based on statistical or structural similarities (i.e., correlation, mutual information), Euclidean distance, or even in a random way (Zhu *et al.*, 2021).
 - The third main step of this algorithm comprises ranking each pair of features based on how close or similar they are. As a result of this ranking, the features would be placed in the image (Zhu *et al.*, 2021).
 - The fourth and final main step of this algorithm entails the optimization process to minimize the differences (i.e., the goal is to ensure that features that are similar are placed in neighbouring pixels on the image) between the following rankings (Zhu *et al.*, 2021):
 - The ranking of similarities between the features.
 - The ranking of how close or far apart the pixels are on the image grid.
- Despite the information described in the last three main steps (i.e. main step two, three and four), in the context of this project the assignment of features to specific pixel position of the image is completely random but the same for all images (i.e. consistency across all images), therefore there is no relationship between the features and their spatial positions in the image (i.e., similar features might be far apart in the image).
- Each row of the tabular dataset is going to be transformed into an image.
- A brief contrast of the algorithm “Image Generator for Tabular Data” with other algorithms that also transform tabular datasets into images is outlined in the subsequent points (Zhu *et al.*, 2021):
 - The algorithm “Image Generator for Tabular Data” doesn't require domain knowledge unlike the algorithm “OmicsMapNet”.
 - The algorithm “Image Generator for Tabular Data” delivers more compact image representations (i.e., a single pixel represents a single feature) in comparison to the algorithm “DeepInsight.” This means that less memory and time are required to train the convolutional neural network compared to other algorithms that generate larger images with unused pixels.

- The algorithm “Image Generator for Tabular Data” preserves the surrounding feature structure much better compared to the algorithm “REFINED.”
- Every image produced after using this algorithm has a size of 224 x 224 pixels; therefore, this is the image size of every input image of the convolutional neural network (see Figure 5 and Figure 6).

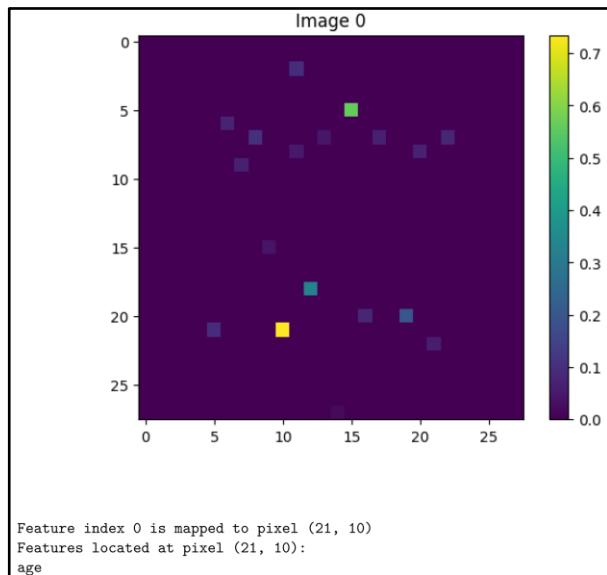


Figure 5. Instance of the training set transformed into an image

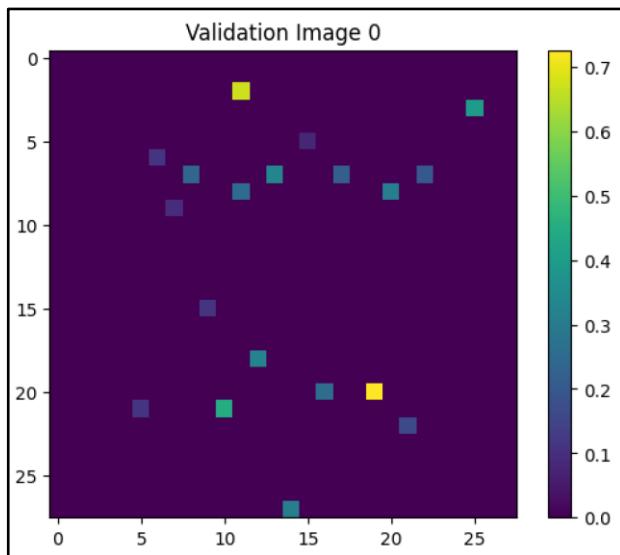


Figure 6. Instance of the validation set transformed into an image

4.8.3 Convolutional Neural Network – Training Phase, Validation Phase and Test Phase

- Data used: stratified training set, stratified validation set, stratified test set (all features)

The learning algorithm applied was a vanilla convolutional neural network. This type of algorithm had the following key components:

- Input Layer: this layer receives the images as a tensor represented with the width, height.
- Convolutional Layers: these layers apply convolutional operations to the input with the aim to extract several features or patterns of the input images. "Kernels" (3x3 pixels), and "padding" (1 extra pixel) were applied to control the size of the output of the convolution process.
- Activation Function: ReLU Function or also called Rectified Linear Unit Function was the chosen activation function to help the network learn complex patterns by introducing non-linearity. It was preferred because it avoids vanishing gradient problems, as its gradients are either 0 or 1.
- Pooling Layer: "MaxPool2d" was the preferred choice of layers over average pooling to extract dominant features from input and reduce spatial dimensions.
- Fully Connected Layers: After the convolutional phase (which uses several sets of convolutional layers together with "ReLU" functions and "MaxPool2d" layers), the output is flattened and fed as input to several fully connected layers. These fully connected layers combine the final extracted features from the convolutional phase.
- Output Layer: This final layer (located at the end of the Fully Connected Layers) produces a single value for the regression task.
- Loss Function: The "MSELoss" (i.e., the Mean Square Error) was the function chosen to measure the difference between the predicted output and the true values during the training and validation phase.
- Optimizer: Adaptive Moment Estimation (also called "Adam Optimizer") was the optimization technique for gradient descent (i.e., to update the network's weights and minimize the value of the loss function " MSELoss").

Some key details about the vanilla convolutional neural network applied in this project are described in the following lines:

- Architecture:
 - o It consists of two convolutional layers (conv1 and conv2) followed by max-pooling layers (maxpool1 and maxpool2).
 - o After the convolutional layers, there's a flattening operation to convert the 2D feature maps into a 1D vector.
 - o Two fully connected layers (fc1 and fc2) are used for the regression output.

- Layers and Parameters:
 - o conv1: A 2D convolutional layer with 1 input channel, 32 output channels, and a kernel size of 3x3. It uses padding to maintain spatial dimensions.
 - o relu1: the ReLU activation function was applied after conv1.
 - o maxpool1: Max-pooling layer with a 2x2 kernel size.
 - o conv2: Another 2D convolutional layer with 32 input channels (from conv1), 64 output channels, and a 3x3 kernel size.
 - o relu2: the ReLU activation function as applied after conv2.
 - o maxpool2: Max-pooling layer with a 2x2 kernel size.
 - o flatten: Flattening layer to convert the 64 feature maps (after maxpool2) into a 1D vector.
 - o fc1: Fully connected layer with 6477 input features (from the flattened output) and 128 output features.
 - o relu3: ReLU activation function applied after fc1.
 - o fc2: Final fully connected layer with 128 input features and 1 output feature for regression.
- Hyperparameters:
 - o Learning rate: 0.001
 - o Batch size: 32
 - o Number of epochs: 20
- Training:
 - o Optimizer: Adam (Adaptive Moment Estimation Algorithm)

The summary of the key outputs of the convolutional neural network in the training set, validation set, and test set is shown in the Table 9 and Figure 7.

Table 9. Convolutional Neural Network performance

Phase	Model	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Mean Square Error (MSE)	Features
Training (Dataset: stratified training dataset)	Convolutional Neural Network	6.43	7.66	58.66	"age", "sex", "test_time", "Jitter (%)", "Jitter (Abs)", "Jitter: RAP", "Jitter: PPQ5", "Jitter: DDP",
		7.17	8.70	75.66	

(Dataset: stratified validation dataset)					"Shimmer", "Shimmer(dB)", "Shimmer: APQ3", "Shimmer: APQ5", "Shimmer: APQ11", "Shimmer: DDA", "NHR", "HNR", "RPDE", "DFA", 'PPE'
Test (Dataset: stratified test dataset)		6.52	7.64	58.43	

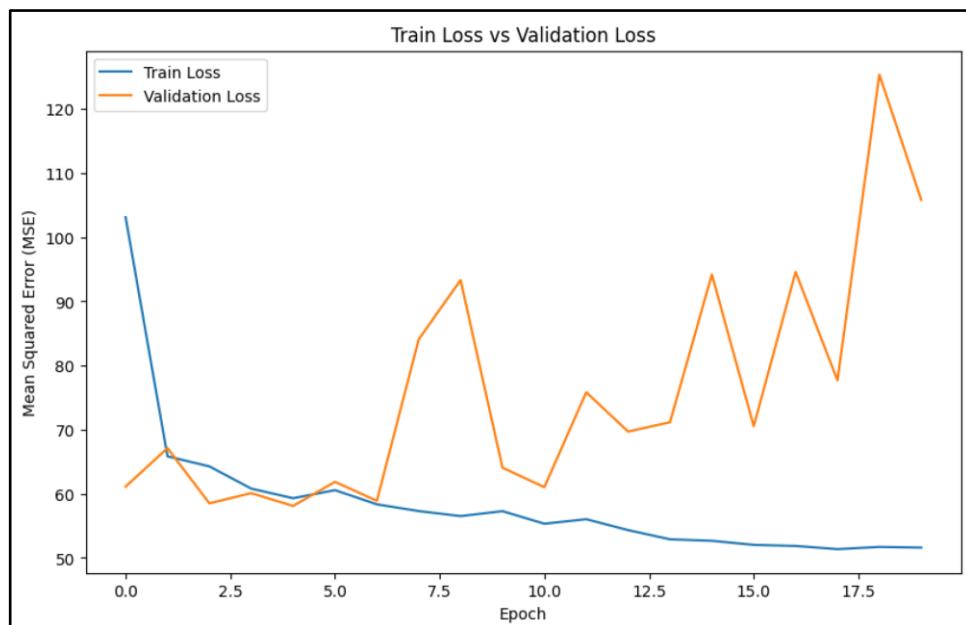


Figure 7. Train Loss vs Validation Loss, CNN

4.8.4 Convolutional Neural Network – Analysis and Evaluation of the results

The difference between the Mean Absolute Error of the convolutional neural network using the stratified training set ($MAE = 6.43$) and the Mean Absolute Error of the trained convolutional neural network in the stratified test set ($MAE = 6.52$) is significantly small (see Table 9). Therefore, it is clear that the capabilities of the convolutional neural network in the context of prediction performance. This is probably due to the fact that the convolutional neural network is well suited to capture intricate insights or patterns of the variables in the dataset.

Furthermore, in Figure 7, both training and validation loss decrease initially (i.e., from epoch number 0 to epoch number 5), suggesting that the model is learning quite well from the training data; however, after epoch number 5, the validation loss starts fluctuating and trending upwards, providing evidence of overfitting patterns. According to the Mean Absolute Error of the Boosted Trees model using the stratified training set ($MAE = 1.21$) and the stratified test set ($MAE = 1.23$) (see Table 8), it is demonstrable that the convolutional neural network was not able to surpass the performance of this simple, fast, efficient, and more interpretable machine learning model.

4.9 Boosted Trees / Convolutional Neural Network - Model Interpretability and Explainability

The concept of Shapley Values was derived from cooperative game theory. Specifically, the Shapley Values method quantifies how much each independent variable contributes to the difference between the actual prediction and the average prediction (Gopinath, 2021). Some interesting properties that made this method the chosen option in this project for both the Boosted Trees and the Convolutional Neural Network were the following:

- It is model-agnostic.
- It can provide local and global interpretability.
- It can handle feature interactions:

4.9.1 Boosted Trees - Model Interpretability and Explainability

The technique of Shapley Values, to be specific the method of Mean Shap Values, which was applied to the Boosted Tree model, was extremely useful in gaining a deep insight into the feature importance in this specific model (see Figure 8 and Figure 9).

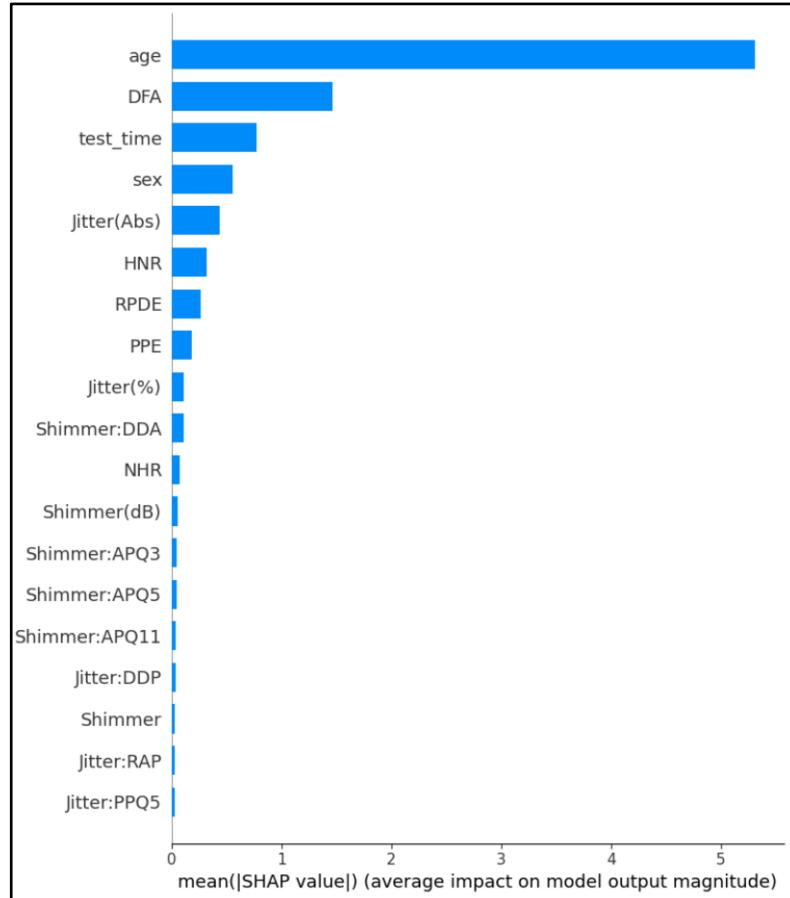


Figure 8. Mean SHAP Value (Average impact on model output), Boosted Trees

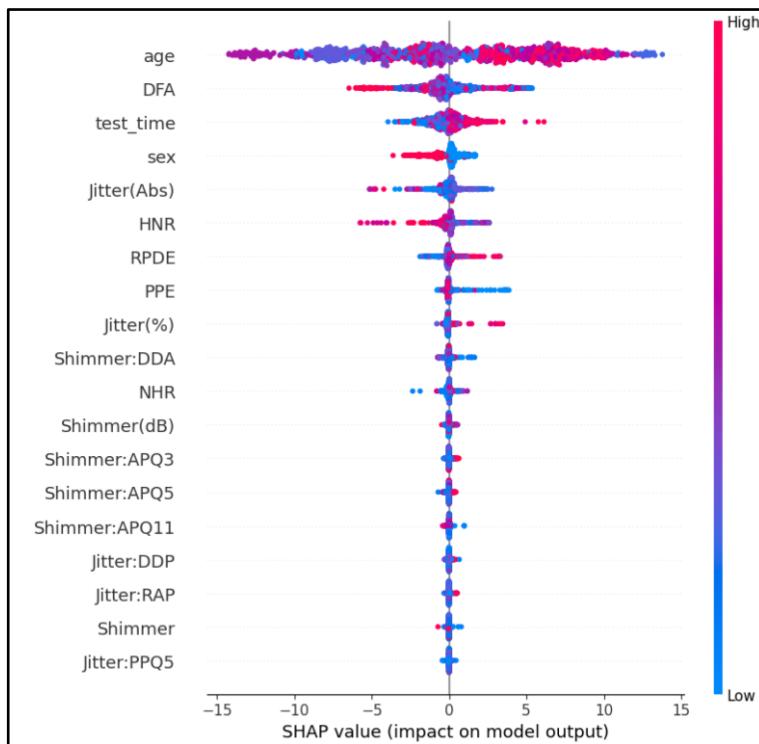


Figure 9. SHAP Value (Impact on model output) – Global Impact, Boosted Trees

4.9.2 Convolutional Neural Network - Model Interpretability and Explainability

The technique of Shapley Values, to be specific the method of Mean SHAP Values, which was applied to the convolutional neural network, was extremely useful to gain a deep insight into the importance of the pixels in the image, which were previously the features of the original dataset. Strictly speaking, Shapley Values display how much each pixel contributes to the predictions of the convolutional neural network (see Figure 10).

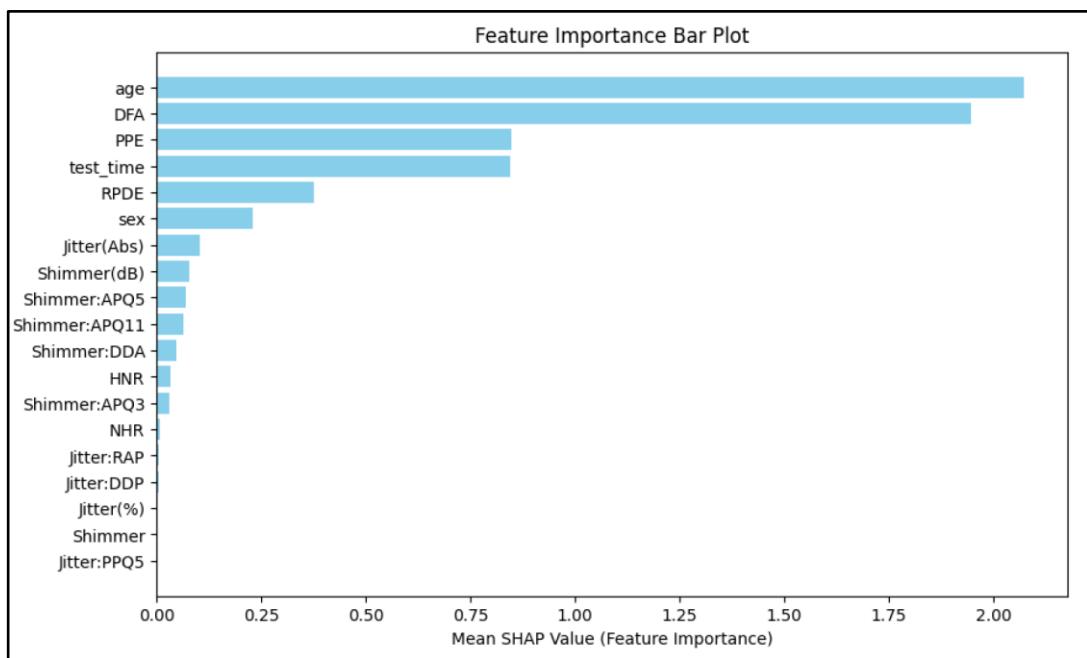


Figure 10. Mean SHAP Value (Feature Importance) – Global Impact, CNN

In the context of local Shapley Values and feature importance, the boosted tree models handled and captured hierarchical feature importance (i.e., they tell the importance of the features that belong to the original tabular dataset), meanwhile, the convolutional neural networks handle and capture spatial hierarchies of feature importance in the images (i.e., pixels or clusters of pixels that could be scattered across the image or not). As a result, the Shapley Value for a convolutional neural network might be more distributed across different regions of the image (i.e., it tells the importance of image regions), making it computationally expensive, hard to interpret, and sometimes difficult to link to the individual importance of the original features of the dataset (see Figure 11).

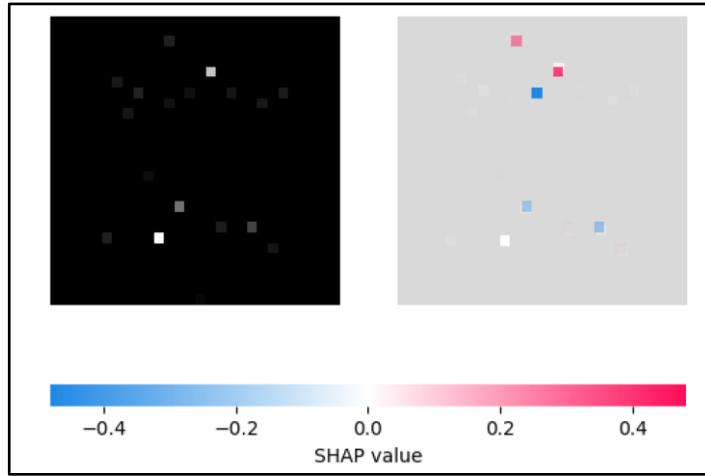


Figure 11. SHAP Value – Local Impact, CNN

The spatial nature of the images generated may align with the relationships in the original tabular dataset because the Shapley Value method (to be specific, the Mean Shapley Value) shows that eight (8) of the top ten (10) variables (see features highlighted in green in the Table 10¹) are in common in both models for the top ten (10) positions.

By comparing the Boosted Trees and the Convolutional Neural Networks trained on different representations of the same dataset (i.e., tabular data and images), the feature importance ranking according to the Shapley Value method (to be specific, the Mean Shapley Value) shows that eight (8) of the top ten (10) variables (see features highlighted in green in the Table 10¹) are common in both models for the top ten (10) positions. This means that the dataset used in this project has properties that can be captured by the deep learning model (i.e., the Convolutional Neural Network) and the machine learning model (Boosted Trees).

Table 10. Top variables according to SHAP Value for Boosted Trees and CNN

N°	Boosted Trees Model Mean Shapley Value	Convolutional Neural Network Mean Shapley Value
1	age	age
2	DFA	DFA
3	test_time	PPE
4	sex	test_time

¹ The variables listed in the Table 10 are sorted from most important according to SHAP values (position N°1) to less important (position N°19).

5	Jitter (Abs)	RPDE
6	HNR	sex
7	RPDE	Jitter (Abs)
8	PPE	Shimmer(dB)
9	Jitter (%)	Shimmer: APQ5
10	Shimmer: DDA	Shimmer: APQ11
11	NHR	Shimmer: DDA
12	Shimmer(dB)	HNR
13	Shimmer: APQ3	Shimmer: APQ3
14	Shimmer: APQ5	NHR
15	Shimmer: APQ11	Jitter: RAP
16	Jitter: DDP	Jitter: DDP
17	Shimmer	Jitter (%)
18	Jitter: RAP	Shimmer
19	Jitter: PPQ5	Jitter: PPQ5

4.10 Boosted Tree Model – Feature Selection

The performance in the regression task based on metrics including Mean Absolute Error, Root Mean Square Error, and Mean Square Error of the Boosted Trees model using all the features is better than the performance of the Convolutional Neural Network with the same number of features (see Section 4.7 and Section 4.8); therefore, in Sections 4.10, 4.11, and 4.12 of this project, we are going to go further in the model development process to accomplish the goal of providing healthcare professionals with an effective and easy-to-use technological tool based on machine learning and explainable artificial intelligence that sheds light on the most important features related to speech measurements for understanding Parkinson's disease.

4.10.1 Feature Selection - Description

Considering the great number of independent variables that are part of the dataset, which consists of twenty independent variables in total, the techniques described in the following lines were applied with the aim of minimizing the number of features, reducing the model's intricacy, and decreasing the levels of noise in the data. Therefore, the model will show an improvement in its generalization capabilities and performance, and consequently moderate its computational complexity.

4.10.2 Feature Selection - Principal Component Analysis

- Data used: stratified training set, stratified validation set, stratified test set (all features)

Despite the fact that the Principal Component Analysis technique does not work properly as an automatic feature selector because its main goal is to reduce the dimensionality of the feature space, which in our case is a feature space of 19 dimensions, by creating new variables called principal components, the Principal Component Analysis technique was applied through the Principal Component Analysis Loading Plot (see Figure 12) to spot positive correlations among the independent features (similar loading values of the features), negative correlations among the features (one positive and one negative loading value on the same principal component), or uncorrelated variables (variables that form a right angle with the plot origin). For instance, see Table 11:

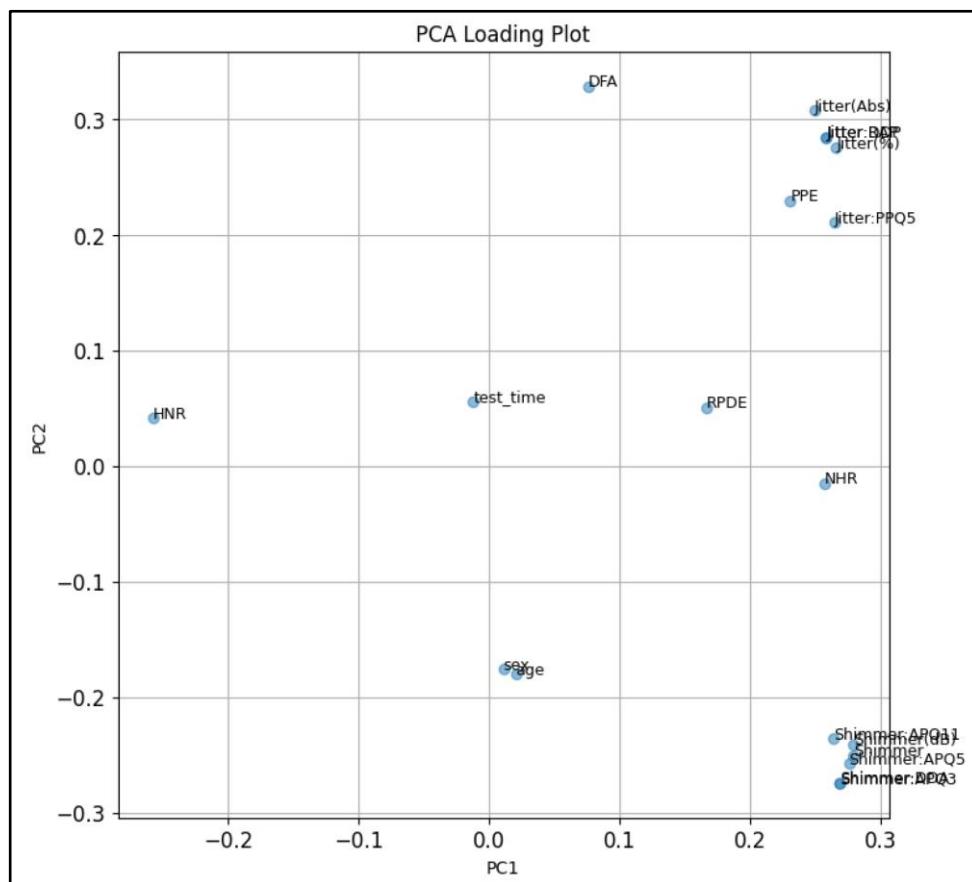


Figure 12. Principal Component Analysis Loading Plot

Table 11. Some PCA Loading correlations

Variables	PCA Loading Correlation
HNR, DFA	Likely to be uncorrelated
HNR, age	Likely to be uncorrelated
HNR, PPE	Negative correlation
DFA, age	Negative correlation
age, sex	Positive correlation

Furthermore, the information about correlations derived from the Principal Component Analysis Loading Plot is very useful at the moment to analyse the outcomes of the Explainable Artificial Intelligence technique called Shapley Values, as this technique tends to be sensitive to correlated features, making it difficult to disentangle the individual contributions of each feature to the output of the model.

4.10.3 Feature Selection - Select K-Best

- Data used: stratified training set, stratified validation set, stratified test set (all features)

This feature selection technique was useful to identify the most relevant independent variables to the dependent variable based on a score from a scoring function. In this project, the scoring function to compute linear correlations between each feature and the target variable is “f_regression” because it was suitable for this regression model, and it was less computationally expensive than other scoring functions. It is essential to note that this method does not capture feature interactions at the moment of computing the scores, as it is only able to capture the individual interactions of each independent variable and the target variable.

Several values of the user-specified parameter “k value” were tried ($k = 2, 4, 6, 8, 10, 12, 14, 16, 18, 20$), and the Boosted Trees model’s performance was evaluated using several error metrics such as Mean Absolute Error, Root Mean Square Error, and Mean Square Error to select the “k value” that provided a good balance between performance and model complexity determined by the number of variables (see Figure 13). After comparing the error metrics listed previously for the several Boosted Trees models, the best “k value” is $k = 4$ (i.e., for the Select K-Best method, the top four produce the lowest value in the error metrics).

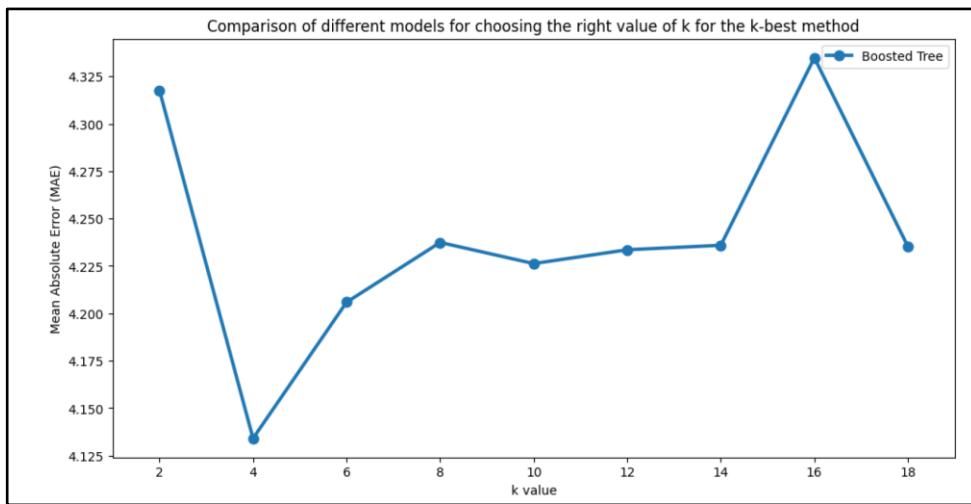


Figure 13. Comparison of different models for choosing the right value of k for the k-Best Method

4.10.4 Feature Selection - Elimination with Cross Validation

- Data used: stratified training set, stratified validation set, stratified test set (all features)

This feature selection technique chose the best features using a method that started with the full set of features and iteratively removed the least important features based on model performance metrics such as Mean Absolute Error, Root Mean Square Error, and Mean Square Error. It will use a machine learning model to rank the importance of features. The least important features are removed in each iteration, and the model's performance is evaluated using cross-validation. This process continues until the optimal number of features is found.

4.10.5 Feature Selection - Sequential Feature Selector

- Data used: stratified training set, stratified validation set, stratified test set (all features)

This feature selection technique chose the best features using a method that iteratively adds or removes independent variables one by one based on a performance criterion of the model, including Mean Absolute Error, Root Mean Square Error, and Mean Square Error. It used a machine learning model (in this case, a linear regression through its coefficients because it was faster and computationally inexpensive compared to other kinds of estimators) to evaluate the performance of the model with different feature subsets and select the subset that optimizes the performance metric.

At the end of the feature selection process, the variables selected were shown in Table 12.

Table 12. Summary of the Feature Selection Process

Feature selection Technique	Number of Features (Before “feature selection” → After “feature selection”)	Feature name
K-best	19 → 4	age
		HNR
		DFA
		PPE
Feature Elimination with Cross Validation	19 → 1	Jitter (RAP)
Sequential Feature Selector	19 → 3	age
		Jitter (Abs)
		DFA

4.11 Boosted Tree Model (Improved Version) – Training Phase, Validation Phase and Test Phase

- Data used: stratified training set, stratified validation set, stratified test set (selected features)

The boosted trees model for a regression task developed in Section 4.7 is clearly overfitting according to the Coefficient of Determination during the training phase ($R^2 = 0.96$) and test phase ($R^2 = 0.90$). Therefore, a Boosted Tree Model (Improved Version) was developed in Sections 4.10, 4.11, and 4.12 with the goal of having a more interpretable regression model and better generalization capabilities. The training, validation, and test phases using the Boosted Trees model (Improved Version) considered as a base the pipelines developed in Section 4.5, Feature Engineering. The summary of the key outputs of the Boosted Trees model in the training set, validation set, and test set is shown in Table 13

Table 13. Boosted Trees (Improved Version) performance

Phase	Model	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Mean Square Error (MSE)	Coefficient of Determination (R^2)	Features
Training (Dataset: stratified training dataset)	Boosted Trees [max_depth: 5]	2.00	2.72	7.42	0.88	
Validation Grid Search Cross validation (Dataset: stratified validation dataset)	Best Model ↓ Boosted Trees [max_depth: 10, n_estimators: 110]	7.10	-	-	-	"age", "Jitter(Abs)", "Jitter:RAP", 'HNR', 'DFA', 'PPE'
Test (Dataset: stratified test dataset)	Best Model ↓ Boosted Trees [max_depth: 10, n_estimators: 110]	2.31	3.52	12.41	0.81	

4.12 Boosted Tree Model (Improved Version) – Analysis and Evaluation of the results

The difference between the Mean Absolute Error (MAE = 1.23) and Coefficient of Determination ($R^2 = 0.90$) of the Boosted Trees model using the stratified test set (see Table 8) with Mean Absolute Error (MAE = 2.31) and Coefficient of Determination ($R^2 = 0.81$) of the Boosted Trees model (Improved Version) using the stratified test set (see

Table 13) shows an enhancement in the generalization capabilities of the model. In other words, the value of the Mean Absolute Error exhibits a moderate increase, and the Coefficient of Determination displays a modest decrease of around 11%.

The relevance of the performance of the Boosted Tree Model (Improved Version) lies in the following points:

- The Coefficient of Determination ($R^2 = 0.81$) of the Boosted Trees model (Improved Version) still quantifies in a reasonable manner the proportion of variance in the target variable (i.e., the variable "motor_UPDRS") that is predictable from the independent variables (i.e., the variables "age," "Jitter (Abs)," "Jitter: RAP," "HNR," "DFA," "PPE").
- The Boosted Trees model with a Coefficient of Determination ($R^2 = 0.90$) was improved because it may have a poor predictive performance on new data or scenarios such as:
 - o Subjects whose language is distinct than English in the voice recordings.
 - o Subject under the effects of medication at the moment of the voice recordings.
 - o Sustained phonation of vowels, words or phrases rather than "ahhh..."
 - o A diverse type of equipment for recording the voices
 - o Subjects in the whole spectrum of Parkinson disease instead of just being in the early-stage of Parkinson's disease.

4.13 Boosted Tree Model (Improved Version - Model Interpretability and Explainability)

- The technique of Shapley Values, to be specific the method of Mean Shap Values, which was applied to the Boosted Tree model, was extremely useful to gain a deep insight into the feature importance in this specific model (see Figure 14 and Figure 15).

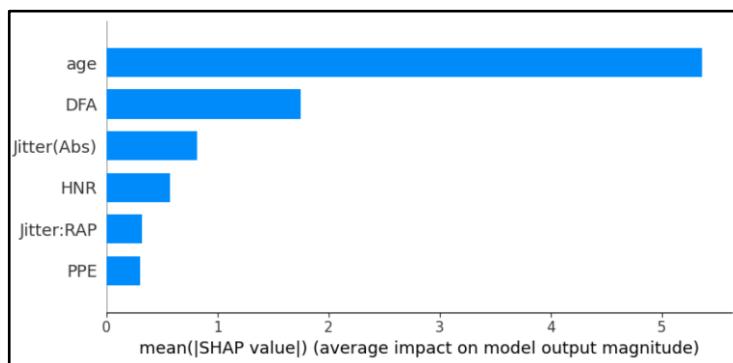


Figure 14. Mean SHAP Value (Average impact on model output) – Global Impact, Boosted Trees

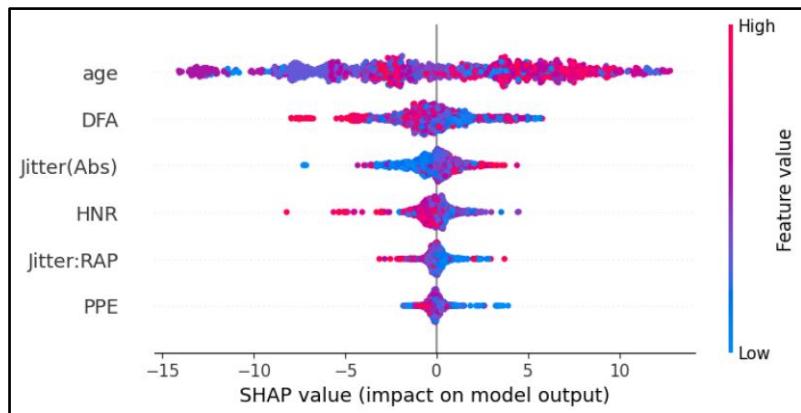


Figure 15. SHAP Value (Impact on model output) – Global Impact, Boosted Trees

4.14 Key Insights of the assessment and evaluation of the results

- The interpretation of the results of this project was done taking into account several factors related to the nature of the input data, features, model architecture, interpretability, and performance.
- Possibly, it would be hard to interpret how the convolutional neural network handles the dataset (i.e., pixel data from images that originally were a tabular dataset) compared to the way the Boosted Tree model worked on the tabular dataset. This means that the input mismatch in the models' inputs used in this project may have made the interpretation and comparison of the results difficult.
- The feature space in which the Boosted Trees models and the convolutional neural network were working was not the same; therefore, the comparison could be considered problematic:
 - o The boosted trees models handled and interpreted features in their original format.
 - o The convolutional neural network handled and interpreted features in their image format (i.e., transformed format), which may introduce spatial interactions that were not present in the original dataset. Consequently, the application of Shapley Value methods through a convolutional neural network may focus on pixels or regions of pixels in

the image that are important for its predictions, but may not be associated with the original features.

- By almost any metric, including the Mean Absolute Error, Root Mean Square Error, and Mean Square Error, the performance shown by the Convolutional Neural Network surpassed the performance of the Boosted Trees model and the Boosted Trees model (Improved Version). The interpretation of these outcomes leads to the following considerations:
 - o The dataset of this project was originally a structured dataset (i.e., a tabular dataset), and the transformation into images, which in this case was through the novel algorithm called “Image Generator for Tabular Data,” could potentially introduce some complexity, loss of information, or noise degrading the performance of the convolutional neural network. Moreover, this could be an indication of the suitability of the Boosted Trees models for tabular data.
 - o Probably, the transformation from tabular data to images using the algorithm “Image Generator for Tabular Data” was not adequate enough; therefore, the convolutional neural network did not discover non-trivial patterns, spatial relationships, or spatial hierarchies that could put it ahead of the Boosted Trees models in performance metrics.
 - o On the face of a convolutional neural network that poorly performed versus the Boosted Trees models, it is reasonable to indicate that maybe the image transformation using the algorithm “Image Generator for Tabular Data” was not highly effective in preserving meaningful feature interactions.
- Considering the poor performance of the convolutional neural network, it is highly likely that it required a deep tuning process (e.g., learning rate, batch size, number of layers, number of pooling layers, functions) to make sense of the transformed tabular data, contrasting with the lightweight hyperparameter tuning process of the Boosted Trees models.

The convolutional neural network was computationally intensive compared to the boosted trees models. In addition, this kind of model (i.e., convolutional neural network) tends to require a lot of data (i.e., a data-hungry model) to work well. The previous considerations described in this paragraph make feasible the use of boosted trees models as the model to face or approach these limitations.

- The features ["age", "Jitter (Abs)", "Jitter: RAP", "HNR", "DFA", "PPE"] used in the development of the Boosted Trees model (Improver Version) were sufficient to achieve good performance compared to using the whole dataset used in the Boosted Trees and Convolutional Neural Network; therefore, the evidence suggests that there were some irrelevant or redundant features that were discarded by the feature selection methods such as Select K-Best, Feature Elimination with Cross Validation, and Sequential Feature Selector.

CHAPTER 5. - FURTHER WORK

5.1 Introduction

The purpose of this chapter is to comment on the problems that arose during several stages of the dissertation project. Moreover, some ideas for future study will be listed and addressed in this section.

5.2 Chapter 3. Data – Challenges

Regardless of the effort and time invested in looking at several websites and platforms for more datasets in the context of biomedical speech measurements of subjects with Parkinson's Disease to train, validate, and test the models, obtaining more data was unachievable.

5.3 Chapter 4. Data – Design, Implementation and Results - Challenges

During several days and in different stages of the model design, building, and implementation, some issues with the cloud-based platform Google Collaboratory (see Figure 16 and Figure 17) were encountered and were not resolved despite many attempts and approaches. This situation had a negative impact on the number and variety of experiments performed in this dissertation project.

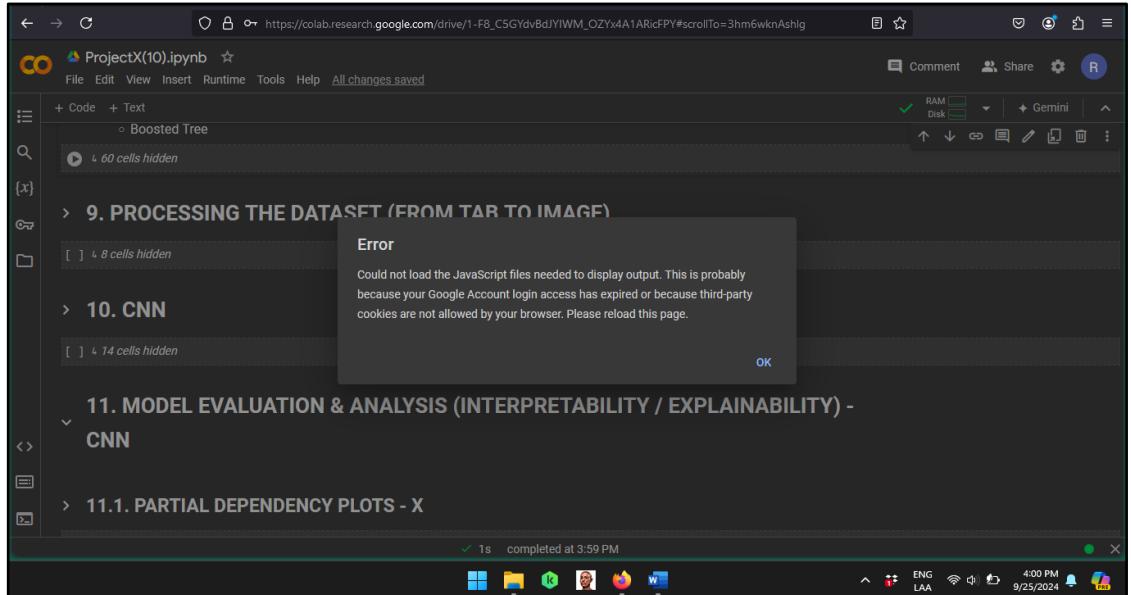


Figure 16. Error message in the Google Collaboratory platform (September 25, 2024)

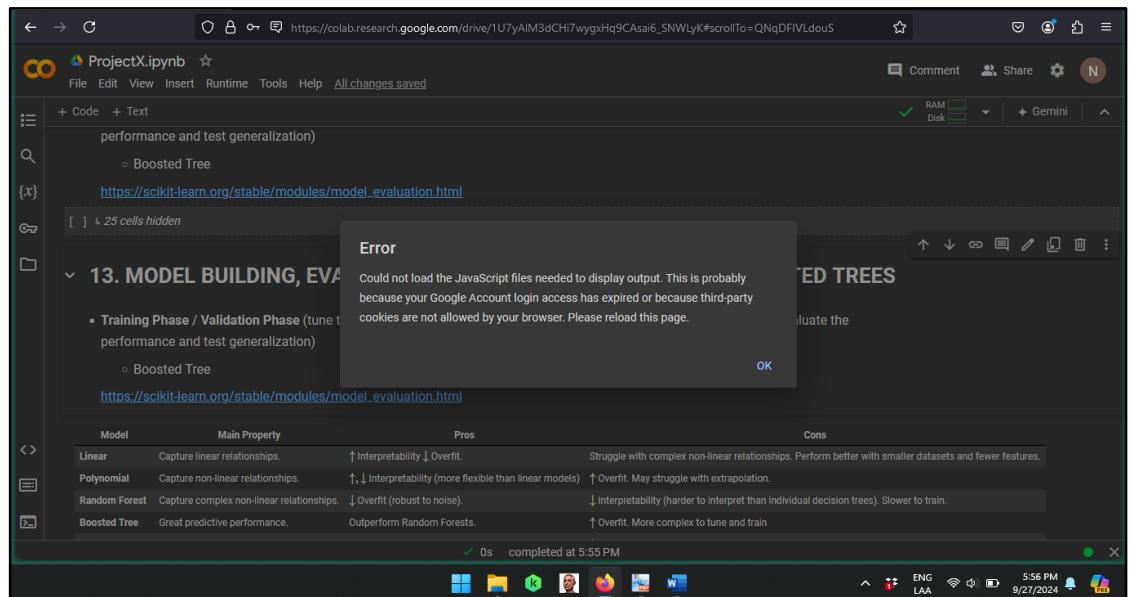


Figure 17. Error message in the Google Collaboratory platform (September 27, 2024)

5.4 Future ideas

Based on the experience gained developing this project, some recommendations and/or advice for the success of future endeavours in the same research field are the following:

- Considering the limiting factor of the amount of data publicly available, it could be relevant to consider the application of data augmentation techniques to the available dataset. This approach could be useful in the context of small datasets jeopardizing the generalization capabilities of the model. In addition, more data means fewer models that use multiple voice recordings of the same subject in the training, validation, and test sets.
- In the context of splitting the data into training, validation, and test sets, other approaches rather than stratified splitting should be applied to assess the performance of the model under several conditions
- In the right circumstances and conditions, the use of pretrained convolutional neural networks could improve the amount and complexity of patterns captured from the data and, as a consequence, have a positive impact on the performance of the model. In addition, pre-trained convolutional neural networks could leverage the knowledge from other dataset domains.
- The implementation of other approaches for transforming the tabular dataset into images (for instance, grouping or clustering pixels that are highly correlated during the transformation of tabular data into images) could be helpful to conduct more experiments about the feasibility of convolutional neural networks in this topic.
- The impact of the fine-tuning process of the Boosted Trees model and the convolutional neural networks, considering the task (i.e., regression task) and the data conditions, should not be neglected.
- Use graphics processing units, also called GPUs, to accelerate all the phases of the development, experimentation, and implementation process of a convolutional neural network.

CHAPTER 6. - CONCLUSIONS

Basically, the work undertaken consisted of the design, build, execution, assessment, and comparison of the performance metrics and interpretability for a regression task of a Boosted Tree model and a convolutional neural network in a dataset of biomedical speech measurements of subjects with early-stage Parkinson's Disease. The Boosted Tree model worked with the numeric dataset in tabular format, while the convolutional neural network used the same dataset but transformed it into images. Therefore, the main aim of the project to provide healthcare professionals with an effective and easy-to-use technological tool based on machine learning or deep learning and explainable artificial intelligence that sheds light on and describes the most important features related to speech measurements for understanding Parkinson's disease was achieved.

Furthermore, the three objectives set in the Chapter 1. Introduction to help accomplish this main goal were fulfilled as follows:

- Literature Review – See Chapter 2
- Exploratory data analysis and data processing – See Chapter 3 and Chapter 4
- Design, build, and run a machine learning and deep learning model – See Chapter 4

The Boosted Trees model scored a Mean Absolute Error of 1.21 for the training phase and 1.23 for the testing phase on the metric tabular dataset; meanwhile, the convolutional neural network achieved a Mean Absolute Error of 6.43 for the training phase and 6.52 for the testing phase on the image dataset. In addition, both models display 8 features in common in the top 10 features that have the most impact on model prediction according to the Mean Shapley Value technique. Based on the analysis of the models' outputs in the context of performance and interpretability, the Boosted Trees model (Improved Version) is the model chosen as the effective and easy-to-use technological tool based on machine learning and explainable artificial intelligence.

Once again, it is extremely important to note the relevance of the concept of interpretability in the field of healthcare and in the decision taken in favour of the Boosted Trees model. The multidimensional relationships and complex interactions between features of the numeric tabular dataset derived from speech measurements of Parkinson's Disease patients were better captured and interpreted by the Boosted Trees models than by using the convolutional neural network in this regression task.

In conclusion, considering the specific experimental conditions and circumstances of this project, the respective transformation of the tabular dataset into images, the interpretability of the model, performance metrics, computational resources, complexity during the fine-tuning process, and time consumption, there is evidence to consider that it is not feasible to use the convolutional neural network as a technique to provide an effective and easy-to-use technological tool based on deep learning and explainable artificial intelligence to shed light on the most important features related to speech measurements for understanding Parkinson's disease.

REFERENCES

- Bárcenas, R., Fuentes-García, R. and Naranjo, L. (2022) 'Mixed kernel SVR addressing Parkinson's progression from voice features', PLOS ONE. Edited by Y. Tang, 17(10), p. e0275721. Available at: <https://doi.org/10.1371/journal.pone.0275721> (Accessed: 18 August 2024).
- Bowen, L. et al. (2014) 'Effects of Parkinson's Disease on Fundamental Frequency Variability in Running Speech', Journal of Medical Speech-Language Pathology, 21(3), pp. 235–244. Available at: https://sites.bu.edu/steplab/files/2015/02/Bowen.Hands_.Pradhan.Stepp_.2014.pdf (Accessed: 1 September 2024).
- Cerri, S., Mus, L. and Blandini, F. (2019) 'Parkinson's Disease in Women and Men: What's the Difference?', Journal of Parkinson's Disease, 9(3), pp. 501–515. Available at: <https://doi.org/10.3233/jpd-191683> (Accessed: 21 September 2024).
- Crispino, P. et al. (2020) 'Gender Differences and Quality of Life in Parkinson's Disease', International Journal of Environmental Research and Public Health, 18(1), p. 198. Available at: <https://doi.org/10.3390/ijerph18010198> (Accessed: 21 July 2024).
- Fernandes, J. et al. (2018) 'Harmonic to Noise Ratio Measurement - Selection of Window and Length', Procedia Computer Science, 138, pp. 280–285. Available at: <https://doi.org/10.1016/j.procs.2018.10.040> (Accessed: 8 August 2024).
- Gibbons, C.H. et al. (2024) 'Skin Biopsy Detection of Phosphorylated α -Synuclein in Patients with Synucleinopathies', JAMA [Preprint]. Available at: <https://doi.org/10.1001/jama.2024.0792> (Accessed: 3 August 2024).
- Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning. MIT Press. Available at: <http://www.deeplearningbook.org> (Accessed: 1 October 2024).
- Gopinath, D. (2021) The Shapley Value for ML Models - Towards Data Science, Medium. Towards Data Science. Available at: <https://towardsdatascience.com/the-shapley-value-for-ml-models-f1100bff78d1?gi=9fd3a2fc6c67> (Accessed: 3 October 2024).

Hecker, P. et al. (2022) 'Voice Analysis for Neurological Disorder Recognition—A Systematic Review and Perspective on Emerging Trends', *Frontiers in Digital Health*, 4. Available at: <https://doi.org/10.3389/fdgth.2022.842301> (Accessed: 3 August 2024).

Ikuma, T. et al. (2022) 'Harmonics-to-noise ratio estimation with deterministically time-varying harmonic model for pathological voice signals', *The Journal of the Acoustical Society of America*, 152(3), pp. 1783–1794. Available at: <https://doi.org/10.1121/10.0014177> (Accessed: 16 August 2024).

Kiran, A., Mounika, M. and Srinivas, B. (2015) 'Statistical Comparison of Jitter and Shimmer Voice Features for Healthy and Parkinson Affected Persons', *Journal of Engineering Sciences*, 6(2). Available at: <https://jespublication.com/upload/2015-V6I22015.pdf> (Accessed: 6 September 2024).

Little, M.A. et al. (2009) 'Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease', *IEEE Transactions on Biomedical Engineering*, 56(4), pp. 1015–1022. Available at: <https://doi.org/10.1109/tbme.2008.2005954> (Accessed: 8 August 2024).

Ma, A., Lau, K.K. and Thyagarajan, D. (2020) 'Voice changes in Parkinson's disease: What are they telling us?', *Journal of Clinical Neuroscience*, 72, pp. 1–7. Available at: <https://doi.org/10.1016/j.jocn.2019.12.029> (Accessed: 8 August 2024).

National Institute of Neurological Disorders and Stroke (2023) Parkinson's Disease, www.ninds.nih.gov. Available at: <https://www.ninds.nih.gov/health-information/disorders/parkinsons-disease> (Accessed: 27 July 2024).

Naveen Kumar, K. (2023)  Parkinson's Disease Detection Dataset  , Kaggle.com. Available at: <https://www.kaggle.com/datasets/naveenkumar20bps1137/parkinsons-disease-detection/data?select=parkinsons.data> (Accessed: 29 June 2024).

NHS (2022) Parkinson's Disease, NHS. NHS. Available at: <https://www.nhs.uk/conditions/parkinsons-disease/> (Accessed: 15 July 2024).

Parkinson's Foundation (2024) Statistics, Parkinson's Foundation. Parkinson's Foundation. Available at: <https://www.parkinson.org/Understanding-Parkinsons/Statistics> (Accessed: 21 September 2024).

Parkinson's UK (2020) Non-motor symptoms of Parkinson's | Parkinson's UK, Parkinsons.org.uk. Available at: <https://www.parkinsons.org.uk/information-and-support/non-motor-symptoms-parkinsons> (Accessed: 6 July 2024).

Pawar, S., Bais, S. and Bhadule, Y. (2024) 'A Comprehensive Review of Parkinson Disease INTRODUCTION', International Journal of Pharmacy and Herbal Technology, 2(1). Available at: <https://ijprdjournal.com/myapp/uploads/50-926-937%20Yash%20S.%20Bhadule.pdf> (Accessed: 1 July 2024).

Suppa, A. et al. (2022) 'Voice in Parkinson's Disease: A Machine Learning Study', Frontiers in Neurology, 13. Available at: <https://doi.org/10.3389/fneur.2022.831428> (Accessed: 22 August 2024).

Tatham, M. and Morton, K. (2006) 'Speech Production: Prosody', Palgrave Macmillan UK eBooks, pp. 121–163. Available at: https://doi.org/10.1057/9780230513969_5 (Accessed: 29 August 2024).

Teixeira, J.P. and Gonçalves, A. (2014) 'Accuracy of Jitter and Shimmer Measurements', Procedia Technology, 16, pp. 1190–1199. Available at: <https://doi.org/10.1016/j.protcy.2014.10.134> (Accessed: 3 September 2024).

The Voice Foundation (2019) Understanding Voice Production, THE VOICE FOUNDATION. Available at: <https://voicefoundation.org/health-science/voice-disorders/anatomy-physiology-of-voice-production/understanding-voice-production/> (Accessed: 19 August 2024).

Tracy, J.M. et al. (2020) 'Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease', Journal of Biomedical Informatics, 104, p. 103362. Available at: <https://doi.org/10.1016/j.jbi.2019.103362> (Accessed: 27 July 2024).

Tsanas, A. et al. (2010) 'Accurate Telemonitoring of Parkinson's Disease Progression by Non-invasive Speech Tests', IEEE Transactions on Biomedical Engineering, 57(4), pp.

884–893. Available at: <https://doi.org/10.1109/tbme.2009.2036000> (Accessed: 3 July 2024).

UCI Machine Learning Repository (no date) archive.ics.uci.edu. Available at: <https://archive.ics.uci.edu/dataset/174/parkinsons> (Accessed: 3 July 2024).

Wang, T.V. and Song, P.C. (2022) 'Neurological Voice Disorders: A Review', International Journal of Head and Neck Surgery, 13(1), pp. 32–40. Available at: <https://doi.org/10.5005/jp-journals-10001-1521> (Accessed: 24 July 2024).

Williamson, G. (2014) Acoustic Measures (Norms), SLT info. Available at: <https://www.sltinfo.com/acoustic-measures-norms/> (Accessed: 16 August 2024).

Zhu, Y. et al. (2021) 'Converting tabular data into images for deep learning with convolutional neural networks', Scientific Reports, 11(1). Available at: <https://doi.org/10.1038/s41598-021-90923-y> (Accessed: 1 August 2024).

APPENDICES

1. Appendix A. - Terms of Reference

7Z10SS Masters Project	NPC	ToR Coversheet
Department of Computing and Mathematics Computing and Digital Technology Postgraduate Programmes Terms of Reference Coversheet		
Student name:	Ramses Moreno De La Cruz	
University I.D.:	23624012	
Academic supervisor:	Dr. Indranath Chatterjee	
External collaborator (optional):		
Project title:	Understanding of Parkinson's Disease (PD) through Explainable Artificial Intelligence (XAI) applied to speech measurements	
Degree title:	MSc Data Science	
Project unit code:	6G7V0007_2324_9F	
Credit rating:	60	
Start date:	21/06/24	
ToR date:	21/06/24	
Intended submission date:	27/09/24	
Signature and date student:	Ramses Moreno De La Cruz 19/06/2024	
Signature and date external collaborator (if involved):		
MMU	1	CMDT

TERMS OF REFERENCE (TOR)

1. PROJECT TITLE

Understanding of Parkinson's Disease (PD) through Explainable Artificial Intelligence (XAI) applied to speech measurements.

2. COURSE-SPECIFIC LEARNING OUTCOMES

The project will allow the development of the following knowledge and skills:

- Choose, customize, and integrate machine learning and deep learning techniques to build more explainable or interpretable model solutions for approaching real-world data science problems in the health field related to Parkinson's Disease (PD).
- Process and analyze data of speech measurements of people with Parkinson's Disease (PD) effectively and efficiently, using a range of suitable tools.

3. AIM

Provide healthcare professionals with an effective and easy-to-use technological tool based on machine learning, deep learning, and explainable artificial intelligence (XAI) that sheds light on the way the models make decisions and describes the most important features related to speech measurements for understanding Parkinson's Disease (PD)

4. PROJECT DESCRIPTION / SPECIFIC OBJECTIVES

Considering the main goal described in the previous section, the main work to undertake can be described in the following specific objectives:

- Conduct a literature survey and review in the following fields: Parkinson Disease, speech measurement, deep learning, machine learning, and explainable artificial intelligence (XAI).
- Data processing of the dataset of people with Parkinson's Disease (PD).
- Design, build, and run a machine learning and deep learning model for a regression task using the dataset of people with Parkinson's Disease (PD). Then, explainable artificial intelligence (XAI) techniques will be applied to explain the regression task and identify the most relevant features for the regression algorithm.

5. EVALUATION PLAN

The evaluation process of the built model will be based on assessing changes in performance metrics.

6. REQUIRED RESOURCES

The most important hardware resources required are as follows:

- Laptop or desktop
- Graphics processing units (GPUs)

7. REFERENCES

Some of the most relevant bibliographic and infographic references are the following:

- Tsanas, A. et al. (2010) 'Accurate Telemonitoring of Parkinson's Disease Progression by Noninvasive Speech Tests', *IEEE Transactions on Biomedical Engineering*, 57(4), pp. 884–893. Available at: <https://doi.org/10.1109/tbme.2009.2036000>.
 - <https://www.nature.com/articles/npre.2009.3920.1.pdf>
- Little, M.A. et al. (2009) 'Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease', *IEEE Transactions on Biomedical Engineering*, 56(4), pp. 1015–1022. Available at: <https://doi.org/10.1109/tbme.2008.2005954>.
 - <https://pubmed.ncbi.nlm.nih.gov/21399744/>
- Data set: biomedical voice measurements
 - <https://www.kaggle.com/datasets/naveenkumar20bps1137/parkinsons-disease-detection>

8. ACTIVITY SCHEDULE

The project schedule was organized using a Gantt, as is described below.

TASK / MILESTONE	START DATE	END DATE	DURATION (DAYS)
Terms of References & ethOS Platform	31-May-2024	20-Jun-2024	20
Literature Review	31-May-2024	30-Jun-2024	30
Data Analysis and Model Development	30-Jun-2024	20-Jul-2024	20
Report Writing	20-Jul-2024	10-Sep-2024	52
Project Submission	10-Sep-2024	27-Sep-2024	17

TASK	May	June			July			August			September		
	31	10	20	30	10	20	30	10	20	30	10	20	27
Terms of References & ethOS Platform													
Literature Review													
Data Analysis and Model Development													
Report Writing													
Project Submission													

START HERE - Basic Information

This form must be completed for all student projects.

Before you proceed

Some activities inherently involve increased risks or approval by external regulatory bodies, so a proportional ethics review is not recommended and a full ethical review may be required.

These may include:

- i. Approval from an external regulatory body (including, but not limited to: NHS (HRA), HMPPS etc.);
- ii. Misleading participants;
- iii. Research without the participants' consent;
- iv. Clinical procedures with participants;
- v. The ingestion or administration of any substance to participants by any means of delivery;
- vi. The use of novel techniques, even where apparently non-invasive, whose safety may be open to question;
- vii. The use of ionising radiation or exposure to radioactive materials;
- viii. Engaging in, witnessing, or monitoring criminal activity;
- ix. Engaging with, or accessing terrorism related materials;
- x. A requirement for security clearance to access participants, data or materials;
- xi. Physical or psychological risk to the participants or researcher;
- xii. The project activity takes place in a country outside of the UK for which there is currently an active travel warning issued by the authorities (see info button);
- xiii. Animals, animal tissue, new or existing human tissue, or biological toxins and agents;
- xiv. The sharing of participant personal data with a third party, regardless of the form under which the data is presented.

If any of these activities are fundamental to your project, please contact your supervisor to determine if a full application is required.

This form must be completed for each research project which you undertake at the University. It must be approved by your supervisor (where relevant) PRIOR to the start of any data collection.

In completing this form, please consult the University's [Research Ethics and Governance standards](#).

A1a Please confirm that you will abide by the University's Research Ethics and Governance standards in relation to this project.

- Yes
 No

A1b Data Protection

The University is responsible for complying with the UK General Data Protection Regulation whenever personal data is processed. Under the Data Protection Policy, all staff and students have a responsibility to comply with the regulation in their day-to-day activities. The first step you can take to understand these responsibilities is to review the [Data Protection in Research guidance pages](#) and complete the University's Mandatory Data Protection Training. Student training is available through Moodle (in the 'Skills Online' section – [please follow this link](#)). To make sure your knowledge is up to date, all staff and students must complete the training every two years. If you have any issues in accessing the data protection training or have any questions about the training, please contact dataprotection@mmu.ac.uk.

Have you reviewed the Data Protection guidance pages and completed the Data Protection Training in the last two years?

- Yes
 No

A2 Are you submitting this application as a learning experience, for a unit which already has ethical approval? (please confirm with your supervisor)

- Yes
 No

A2.1 Approval reference (supplied by your supervisor)**A3 Student details**

Title	First Name	Surname
-------	------------	---------

<input type="text"/>	<input type="text" value="Ramses"/>	<input type="text" value="Moreno De La Cruz"/>
----------------------	-------------------------------------	--

Email	<input type="text" value="RAMSES.MORENO-DE-LA-CRUZ@stu.mmu.ac.uk"/>
-------	---

A3.1 Manchester Metropolitan University ID number**A4 Supervisor**

Title	First Name	Surname
-------	------------	---------

<input type="text" value="Dr."/>	<input type="text" value="Indranath"/>	<input type="text" value="Chatterjee"/>
----------------------------------	--	---

Faculty	<input type="text" value="Science and Engineering"/>
---------	--

Telephone	<input type="text" value="+44 (0)161 247 2000"/>
-----------	--

Email	<input type="text" value="i.chatterjee@mmu.ac.uk"/>
-------	---

A5 Which Faculty is responsible for the project?

Science and Engineering



A6 Course title

MSc Project (6G7V0007_2324_9F)

A7 Project title

Understanding of Parkinson's Disease (PD) through Explainable Artificial Intelligence (XAI) applied to speech measurements

A8 What is the proposed start date of your project?

21/06/2024

A9 When do you expect to complete your project?

27/09/2024

A10 Please describe the overall aims of your project (3-4 sentences). Research questions should also be included here.

Provide healthcare professionals with an effective and easy-to-use technological tool based on machine learning, deep learning, and explainable artificial intelligence (XAI) that sheds light on the way the models make decisions and describes the most important features related to speech measurements for understanding Parkinson's Disease (PD).

Research Question: Is the performance of a deep learning model using the whole tabular dataset of voice measurements transformed into images (i.e., considering all the features with their complex patterns and interactions) better than the performance of a traditional machine learning model (that uses only a few selected features)?

A11 Please describe the research activity

- Plan of work - Phase 1:
 1. Dataset loading
 2. Dataset inspection
 3. Feature Selection Process
 4. Use the tabular dataset (based on the feature selection process) in a supervised machine learning algorithm to execute the regression task in the dataset of people with Parkinson's Disease (PD).
 5. Then apply an explainable artificial intelligence (XAI) method to figure out the most relevant features chosen by the supervised machine learning algorithm.

- Plan of work - Phase 2:
 1. Dataset loading
 2. Dataset inspection
 3. Dataset preprocessing / transformation
 4. Use the image dataset in a deep learning algorithm to execute the regression task in the dataset of people with Parkinson's disease (PD).
 5. Then apply an explainable artificial intelligence (XAI) method to determine the most relevant features chosen by the deep learning algorithm.

A12 Please provide details of the participants you intend to involve (please include information relating to the number involved and their demographics; the inclusion and exclusion criteria)

This item does not apply to this project.

A13 Please upload your project protocol

Type	Document Name	File Name	Documents		
			Version Date	Version	Size
Project Protocol	TOR_Ramses Moreno_V3_s	TOR_Ramses Moreno_V3_s.pdf	19/06/2024	1	277.7 KB

Project Activity

B1 Are there any Health and Safety risks to the researcher and/or participants?

- Yes
 No

B2 Please select any of the following which apply to your project

- Aspects involving human participants (including, but not limited to interviews, questionnaires, images, artefacts and social media data)
- Aspects that the researcher or participants could find embarrassing or emotionally upsetting
- Aspects that include culturally sensitive issues (e.g. age, gender, ethnicity etc.)
- Aspects involving vulnerable groups (e.g. prisoners, pregnant women, children, elderly or disabled people, people experiencing mental health problems, victims of crime etc.), but does not require special approval from external bodies (NHS, security clearance, etc.)
- Project activity which will take place in a country outside of the UK
- None of the above

B2.4 Is this project being undertaken as part of a larger research study for which a Manchester Metropolitan application for ethical approval has already been granted or submitted?

- Yes
 No

Data

F1 How and where will data and documentation be stored?

The data and documentation will be stored in the OneDrive cloud storage (i.e., the Microsoft Cloud Storage) from the Manchester Metropolitan University email account.

F2 Will you be using personal data? Personal data is anything than can be used to identify a living individual, directly or indirectly. Pseudonymised data is still personal data.

- Yes
 No

Insurance

F3 Does your project involve:

- Pregnant persons as participants with procedures other than blood samples being taken from them? (see info button)
 Children aged five or under with procedures other than blood samples being taken from them? (see info button)
 Activities being undertaken by the lead investigator or any other member of the study team in a country outside of the UK as indicated in the info button? If 'Yes', please refer to the 'Travel Insurance' guidance on the info button
 Working with Hepatitis, Human T-Cell Lymphotropic Virus Type iii (HTLV iii), or Lymphadenopathy Associated Virus (LAV) or the mutants, derivatives or variations thereof or Acquired Immune Deficiency Syndrome (AIDS) or any syndrome or condition of a similar kind?
 Working with Transmissible Spongiform Encephalopathy (TSE), Creutzfeldt-Jakob Disease (CJD), variant Creutzfeldt-Jakob Disease (vCJD) or new variant Creutzfeldt-Jakob Disease (nvCJD)?
 Working in hazardous areas or high risk countries? (see info button)
 Working with hazardous substances outside of a controlled environment?
 Working with persons with a history of violence, substance abuse or a criminal record?
 None of the above

Additional Information

G1 Do you have any additional information or comments which have not been covered in this form?

- Yes
 No

G2 Do you have any additional documentation which you want to upload?

- Yes
 No

Signatures

H1 I confirm that all information in this application is accurate and true. I will not start this project until I have received Ethical Approval.

- I confirm

H2 Please notify your supervisor that this application is complete and ready to be submitted by clicking "Request" below. Do not begin your project until you have received confirmation from your supervisor - it is your responsibility to ensure that they do this.

Signed: This form was signed by Dr Indranath Chatterjee (I.Chatterjee@mmu.ac.uk) on 19/06/2024 10:59 AM

H3 Have you been instructed by your supervisor to request a second signature for this application?

- Yes
 No

H4 By signing this application you are confirming that all details included in the form have been completed accurately and truthfully. You are also confirming that you will comply with all relevant UK data protection laws, and that that research data generated by the project will be securely archived in line with requirements specified by the University, unless specific legal, contractual, ethical or regulatory requirements apply.

Signed: This form was signed by Ramses Moreno De La Cruz (RAMSES.MORENO-DE-LA-CRUZ@stu.mmu.ac.uk) on 19/06/2024 2:34 AM

Submission - 68634 : Understanding of Parkinson's Disease (PD) through Explainable Artificial Intelligence (XAI) applied to speech measurements

donotreply@infonetica.net <donotreply@infonetica.net>

Wed 6/19/2024 5:09 AM

To:Ramses Moreno De La Cruz <RAMSES.MORENO-DE-LA-CRUZ@stu.mmu.ac.uk>

Cc:Indranath Chatterjee <I.Chatterjee@mmu.ac.uk>

This email originated from outside of Manchester Met. Do not click links or open attachments unless you recognise the sender and believe the content to be safe. Please contact the IT ServiceDesk if you have any concerns, <https://www.mmu.ac.uk/about-us/professional-services/itd/about/contact>

19/06/2024

Project Title: Understanding of Parkinson's Disease (PD) through Explainable Artificial Intelligence (XAI) applied to speech measurements

EthOS Reference Number: 68634

Ethical Opinion

Dear Ramses Moreno De La Cruz,

The above application was reviewed by Dr. Indranath Chatterjee and on the 19/06/2024, was given a favourable ethical opinion. The approval is in place until six months after the end date recorded in your application documentation (27/09/2024).

Approved Documents

Document Type	File Name	Date	Version
Project Protocol	TOR_Ramses Moreno_V3_s	19/06/2024	1

Conditions of favourable ethical opinion

The favourable ethical opinion is granted with the following conditions

Approval is in place for your UG/PGT project

This approval is only valid for Undergraduate (UG) and Post Graduate Taught (PGT) projects and does not grant approval for any Staff or PGR projects.

Adherence to Manchester Metropolitan University's Policies and procedures

This ethical approval is conditional on adherence to Manchester Metropolitan University's Policies, Procedures, guidance and Standard Operating procedures. These can be found on the Manchester Metropolitan University Research Ethics and Governance webpages.

Amendments

If you wish to make a change to this approved application, you will be required to submit an amendment. Please visit the Manchester Metropolitan University Research Ethics and Governance webpages or contact your Faculty research officer for advice around how to do this.

We wish you every success with your project.

Science and Engineering Research Ethics and Governance Committee

v1.10

For help with this application, please first contact your Faculty Research Officer. Their details can be found [here](#)

2. Appendix B. - All Experimentation Code

Available on request from Dr. Indranath Chatterjee