National College of Ireland

Early Diagnosis of Parkinson's Disease Progression

MSc Research Project
Data Analytics

Dharesh Vadalia
Student ID: x18192076

School of Computing
National College of Ireland

Supervisor: Prof. Christian Horn

National College of Ireland
MSc Project Submission Sheet
School of Computing

Student Name:	Dharesh Vadalia		
Student ID:	X18192076		
Programme:	Master of Science - Data Analytics	**Year:**	2020
Module:	M.Sc. Research Project		
Supervisor:	Prof. Christian Horn		
Submission Due Date:	17/8/2020		
Project Title:	Early diagnosis of Parkinson's Disease Progression		
Word Count:	7642	**Page Count:**	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Dharesh Vadalia
Date:	17th August 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table of Contents

Early Diagnosis of Parkinson's Disease Progression

Dharesh Vadalia
x18192076

Abstract

Parkinson's Disease (PD) is triggered due to the loss of dopaminergic neurons in the substantia nigra, disrupting the neural communication of the central nervous system towards reception and response of motor and cognitive senses of the patient. PD is a progressive neural disorder which worsens with ageing. With no clearly outlined pattern posed in symptoms, it is challenging for medical practitioners to identify the disease in its prodromal stage. Inspired from the cause, this research's objective is to predict the rate of progression based on the baseline assessment of a patient so that an appropriate treatment plan can be designed for that individual patient. Biomarkers responsible for baseline assessment are extracted from multiple pre-clinical assessments designed to capture and scale the motor and cognitive impairments experienced by PD patients during the prodromal stage. The study performs clustering of PD patients into 3 clusters marking the rate of progression based on the captured clinical feature of the patients and performance comparison of 7 different ensembled and neural network-based classification model is conducted in this study. The study aims to assist medical practitioners in early diagnosis of risk PD among patients and adopt an appropriate measure to improve patient's quality of life.

Keywords **–** Parkinson's Disease, Clinical Progression, PPMI, Gaussian Mixture Model, Multi-Class Classification

1 Introduction

Neurological disabilities are caused due to dysfunction of the central and peripheral nervous system in humans. Parkinson's disease is one of the most common neurodegenerative disorder impacting 2-3% of the world's population aged above 65 years (Poewe *et al.*, 2017). PD is a result of a degenerative neurological condition caused due to depletion of dopaminergic neuron in the substantia nigra located at midbrain, disturbing patients control over motor and non-motor senses. During the early stages of PD, patients may experience bradykinesia, resting tremors, anxiety, disturbance in sleep, depression, vocal impairment or fatigue (Faivre *et al.*, 2019). These symptoms worsen gradually, causing permanent damage to the patient's quality of life. Patient at the adult age of 60 and above tend to develop visible disabilities, such as postural instability, the freeze of gait and muscle rigidity. Present-day medical and surgical treatments can aid towards containing the spread of disease and delay the occurrence of symptoms in patients by a few years. However, there is no complete cure for PD.

Assessment to identify the risk of PD progression in its prodromal stage is a challenging process. There is no conclusive assessment technique for PD to date, as progression rate of disease and depiction of posed symptomatic patterns varies significantly among different

patients. Clinical practices most commonly employ Unified Parkinson's Disease Rating Scale (UPDRS) in everyday use for evaluating the condition and severity of PD in patients (Goetz *et al.*, 2008). Apart from this various other assessment scales are designed to evaluate the cognitive and motor impairment in PD patients, such as Montreal Cognitive Assessment (MOCA) (Nasreddine *et al.*, 2005), State-Trait Anxiety Inventory (STAI) (Yang *et al.*, 2019), Hopkins Verbal Learning Test (HVLT) (Kuslansky *et al.*, 2004), Geriatric Depression Scale (GDS) (Meara, Mitchelmore and Hobson, 1999), University of Pennsylvania Smell Identification Test (UPSIT) (Driver-Dunckley *et al.*, 2014) and Rapid Eye Movement Sleep Behaviour Disorder scale (REM-RBD). However, the most effective way to diagnose the progression of PD is by analysing the degenerative state of the critically impacted central nervous system using the imaging technique (Bhat *et al.*, 2018). But this technique comes costly and is inappropriate to be employed for early diagnosis of PD. An alternative to which clinical assessment of motor and non-motor sense is far more cost-effective and suitable for early diagnosis.

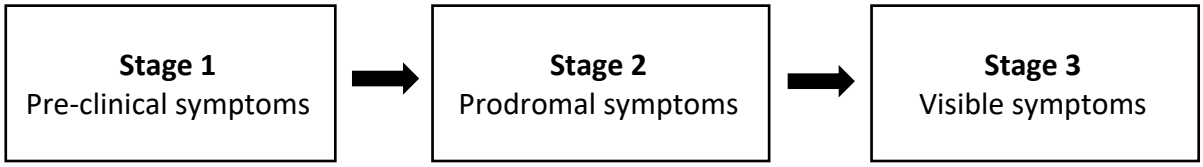

Figure 1: Phases of Parkinson's Diseases progression.

PD patients experience symptoms of PD progression in 3 prominent phases. Phase 1 corresponds to beginning PD due to α-synuclein accumulation in the central nervous system. PD symptoms in this stage are often neglected due to no show or very minimal display of symptoms in patient, leading to misdiagnosis or delay in detecting PD. Phase 2 corresponds to the development of prodromal symptoms (Hindle, 2010). Patients tend to experience mental instability and cognitive disturbances in this stage. PD diagnosis is possible but challenging in this stage using various clinical assessment scales. Phase 3 corresponds to the development of visible motor symptoms. This phase can take about 10-15 years to come into notice after the start of prodromal phase. PD diagnosis is simpler at this stage but is too late to contain the rate of progression and derive a medical monitoring plan.

The proposed research study is based on a novel approach to evaluate early clinical biomarkers of PD to predict the severity and the risk of PD progression in patients. This research aims at understanding various clinical assessment techniques designed by makers of PD biomarkers and identify important clinical features using machine learning approach for predicting the condition of PD. Targeted research tries to answer the question, **how well can combined analysis of clinical biomarkers help identify risk of Parkinson's Disease progression at prodromal stage.** Early identification in case of rapid progression of PD can improve the disease management and monitoring plan for patients.

This paper is formulated into the following sections: Section 2 discusses the brief review of related works conducted in the field of proposed research, highlighting key finding and limitations of applied techniques. Following to which Section 3, outlines the methodology

adopted and design specification for this research work. Section 4, gives a detailed description of end to end steps involved in each stage of project implementation, followed to which model evaluation methods are discussed in Section 5. Discussion over produced results is presented in Section 6 and Section 7 concludes the finding of carried out research work with ideas on future work.

2 Related Work

Much research has been conducted to identify the important biomarkers for diagnosis of PD at its initial stages and it remains a prime area for research in the field of PD. These research methodologies involved various statistical and machine learning approaches to benchmark the primary clinical indicators of PD and identify the pattern of disease progression among different categories of patients. MRI-Scan based analysis is one of the most effective methods for identifying the degenerative state of the central nervous system for diagnosing cases of PD (Bhat *et al.*, 2018). However, this technique is not best suited for identification of PD cases in the prodromal phase as loss of dopamine levels is visually noticeable at a very later stage of PD. To overcome this issue, various evaluation questionnaires with a set of cognitive and motor tests are fabricated to scale the severity of PD among patients. Most popular MDS-UPDRS scale is used as a gold standard for PD assessments (Prashanth and Dutta Roy, 2018). The scale provides a comprehensive set of 63 clinical features divided into 4 parts of the evaluation. Part I encapsulates 13 assessment items for evaluation of non-motor clinical indicator in the patient. Part 2 is concerned with the motor assessment of 13 sets of questions based on evaluation indicators. Part 3 comprises of 33 items which are evaluated by a specialist and lastly Part 4 measure 6 parameters mostly regarding the dosage of levodopa, treatment duration and dyskinesia. UPDRS scale is strongly correlated with another popularly used Hoehn and Yahr (HY) scale. HY scale is highly preferred for gross assessment of PD progression. It scales the risk of progression in 5 stages, ranging from 0 (no presence of PD) to 5 (severe case of PD) (Prashanth and Dutta Roy, 2018). Stage 1 and stage 2 are categorised as early stages where patient experiences and unilateral or bilateral symptoms without any sign of postural imbalance. Stage 3 and stage 4 are considered as moderate stages where the patient is suffering from postural balancing and experiences loss of control over bodily movements. Stage 5 is the extreme phase where the patient is under total medical care.

2.1 Cognitive Assessment based learning

Makers of pre-clinical PD assessment techniques have defined a broad range of indicators to measure the risk of disease progression. Various research studies have been carried out to identify and extract the most contributing list of primary indicators responsible for the diagnosis of PD progression in its initial stage. Research quotes that around 20% of cases of PD are genetically inborn (Bhat *et al.*, 2018). Patients medical history along with the medical history of their family history is considered important for aspect for clinical diagnosis. A research study performed combine learning on 17 features including demographics details and

clinical assessment scores of 553 patients to builds a classification model using Random Forest algorithm for diagnosis of PD and classify progression using HY scale (Soltaninejad, Basu and Cheng, 2019). Biomarkers considered for research include gender, age, history of patients in family and years of education, MDS-UPDRS questionnaire score, MOCA test score, GDS Score and SCOPA-AUT test scores. To calculate the feature importance of all considered features Mean Decrease Impurity (MDI) technique is adopted by the researcher in the following study for feature selection. Following study concluded that demographics data have lower feature importance in diagnosis and determining progression in PD patients. Another research aims at predicting outcomes of cognitive symptoms based on the score of MOCA assessment scale. The author applies Feature Subset Selector algorithms (FSSA) such as Ant colony optimization, Genetic Algorithm and Differential evaluation for selection of most contributing features from MOCA test (Salmanpour *et al.*, 2019). From 10 different prediction models including traditional classification model and neural network model models such as RNN are trained over these features. The comparative evaluation resulted that Least Absolute Shrinkage and Selection Operator Least Angle Regression (LASSOLAR) algorithm delivered best prediction performance.

Another research on stage estimation of PD performed a comparative study on 10 different machine learning models trained over pre-clinical assessment scales from PPMI repository (Prashanth and Dutta Roy, 2018). Models are trained over 59 clinical features of 434 accessed individuals. Evaluation of trained models concluded that cost-sensitive models SVM and AdaBoost algorithm delivered the highest level of classification accuracy. Similar results were produced from an extension of a previous research paper (Challa *et al.*, 2017). In this author introduced Multi perceptron learning and BayesNet model. However, SVM outperformed performance of these neural network model. Also, study inferred that clinical biomarkers such as tremor, bradykinesia, facial expression, and handwriting were marked as the most important feature for PD diagnosis. Similar research by (Tsiouris *et al.*, 2017), quotes that only 10% of patients can be accurately classified to have high risk of progression based on baseline characteristic data captured by PPMI. Also, UPDRS, REM sleep disorder and MOCA assessment tests are highly correlated to each other. Following research adopted a wrapper approach for selection of most optimal feature among all. Classification of progression risk is achieved by Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm. Author (Tsiouris *et al.*, 2018), conducted research using the similar experimental setup for calculating PD progression risk over baseline feature evaluation. Outcomes from research concluded that body bradykinesia, slight speech impairment at baseline are highly correlated. Also, patient classified with a high risk of rapid progression suffer rest tremors for a longer duration. In studied research, author have architected a decision support system (DSS) based on rules extracted from the RIPPER algorithm.

2.2 Motor Assessment based learning

Motor impairment negligible during the early stages of PD, hence mostly ignored. However, early diagnosis of these symptoms can contribute highly towards predicting the rate of progression in patients. In research, the author focused on diagnosing PD stages based on vocal

impairments experienced by the patient. It is one of the most common symptoms posed by PD patients. The study performed a comparison between various ensemble machine learning models and Sparse autoencoder neural network model on vocal data of 188 patients (Xiong and Lu, 2020). For the extraction of features from the vocal data Wolf Search Optimization (WSO) technique is performed. The research concluded LDA with WSO as best performing model for classification of vocal data. Another research in the similar area adopted dimensionality reduction technique PCA to reduce feature dimensions for training these ensemble machine learning models (Celik and Omurca, 2019). Result from the following research highlights Logistic Regression (LR) model as the best performing model. Comparative evaluation of each paper results in WSO as better feature selection technique for vocal data when compared to PCA.

One research focuses on another prime motor symptom of PD patients, tremor. The author tries to diagnose tremor base on sketching pattern of the patient. 11 characteristic metrics are extracted from these sketches using Euclidian distance, Manhattan distance, pixel similarity, sketch time and speed (Bernardo *et al.*, 2019). Based on these metrics three classifier model, SVM, Bayesian classifier and Optimal Path Forest (OPF) are fitted. SVM performance outstands from other classifiers. Another research on the tremor-based diagnosis of PD involved used of digital device to capture the flickering velocity of individuals fingers (Pedrosa *et al.*, 2018). This data was used to train classification models to classify patient between high and low amplitude tremor groups. The research uses Leave-one-out cross-validation (LOOCV) to evaluate model accuracy. Similar research by (Iakovakis *et al.*, 2019) captures touch frequency of patient on touchscreen device over multiple sessions. After each session patients are labelled as HC or PD. Author, architected a neural network model for classification of the patient over labelled data. CNN model designed in following research achieved specificity and sensitivity score of 0.74 and 0.78. Which is reasonable in comparison to other research application.

2.3 Clustering clinical cohorts

The project aims at classifying the case of the patient with respect to their rate of disease progression. Unsupervised clustering techniques can remarkably categorise data points. In a research author aimed at identifying heterogeneity in patients suffering from early stage PD. The author evaluated all markers of PPMI including demographic, motor and cognitive assessment scales. The following research performs a comparison between 3, 4 and 5 number of clusters generated using the K-means clustering technique (Liu *et al.*, 2011). The conducted analysis resulted that most of the variance among PD patient's characteristic based on clinical assessments of PPMI can be explained using 3 clusters. Clusters generated marked patients over rapid progression, normal progression and low progression rate. Another research implements a movie rating system prediction based on Gaussian Mixture Model (GMM). In this study, 4 features extracted for movie rating systems such as rating, number of comments, reviews and view hits are fitted to GMM model to generate clusters. Accuracy of the GMM model is evaluated using a linear regression model. Another research focused on prediction Covid-19 Pandemic using GMM model (Singhal *et al.*, 2020). The author uses a Fourier

decomposition method (FDM) to extract various trends from time-series pandemic data. Decomposed components from FDM are fitted to a GMM model to predict the size of the pandemic. GMM model is not sensitive to exceptions.

Inference's drawn from literature study of related research work conducted in the area of PD progression analysis. Research studies (Bhat *et al.*, 2018) and (Soltaninejad, Basu and Cheng, 2019), guides in the selection of important clinical assessment and its covariates for baseline line evaluation of PD. Correlation between various clinical assessments is drawn from outcomes of (Tsiouris *et al.*, 2017) research. Following papers (Prashanth and Dutta Roy, 2018) and (Challa *et al.*, 2017), gives understanding of the performance of various classification models. SVM algorithm has come out to outperform for PPMI data set in most of the research studies. From (Liu *et al.*, 2011) and (Singhal *et al.*, 2020) performance comparison and application of clustering models acknowledged results GMM to have better-performing edge over K-means clustering algorithm.

3 Research Methodology

Knowledge Discovery in Databases (KDD) and Cross-Industry Standard Process for Data Mining (CRISP-DM) are the two widely used methodologies for research and development in space of machine learning projects. KDD process flow is limited to the evaluation step, unlike CRISP-DM process flow which involves steps for project deployment. As proposed research objective is to evaluate the performance of machine learning models, trained to predict the risk of PD progression, proposed research follows aspects of KDD methodology for implementation.

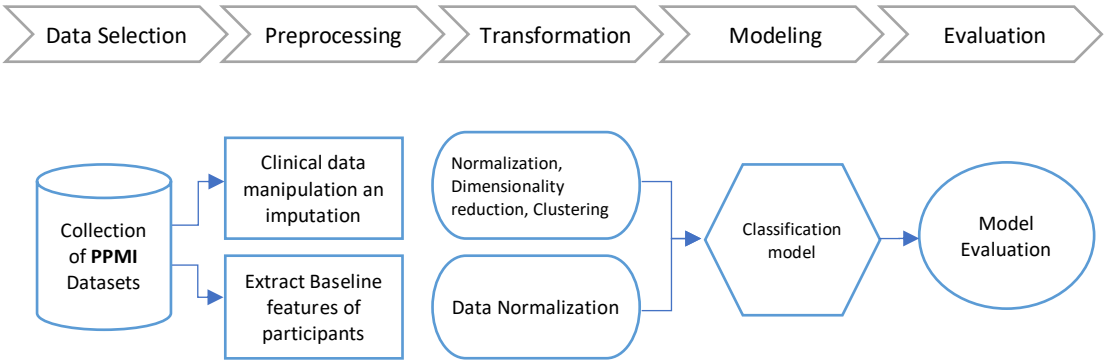

Figure 2: KDD methodology flowchart

Figure 2 illustrates the steps adopted in KDD framework for implementation of this research project. Steps defined in KDD methodology are favourable for complete and accurate implementation of Machine learning models. This section elaborates the process of data gathering and understanding, project design specification and components of the designed model.

3.1 Data Selection

Based on the conducted literature study, it was identified that Parkinson's Progression Markers Initiative (PPMI)[1] an organisation founded by Michael J. Fox Foundation for research in the area of PD, pioneers in the collection of patient's data required to examine PD progression. This comprehensive set of data is maintained in a central public repository available on request for analysis. Dataset offers a complete evaluation of patient clinical condition from baseline to 5 years follow up visits. Captured data holds the clinical test score of various clinical assessment trials conducted on the patient to identify different cognitive and motor developed impairments. Datastore also maintains patient's genetic data and brain MRI-Scan image data for purpose of research. However, in this research clinical assessments, data will be used as an early diagnostic biomarker of PD. Data is collected following standard data acquisition protocols with consent from the patient regarding the use of data for the purpose of research work. PPMI holds data for various clinical assessments for over 1800+ patients, belonging to two prominent categories, suffering from PD and Healthy control (HC).

Category	Assessment	Description
Motor	MDS-UPDRS II	Unified Parkinson's Disease Rating Scale II
	MDS-UPDRS III	Unified Parkinson's Disease Rating Scale III
Cognitive	MDS-UPDRS I	Unified Parkinson's Disease Rating Scale I
	MOCA	Montreal Cognitive Assessment
	STAI	State-Trait Anxiety Inventory
	GDS	Geriatric Depression Scale
	QUIP	Questionnaire for Impulsive-compulsive Disorder in Parkinson's disease
	SCOPA-AUT	Scale for Outcomes in Parkinson's Disease – Autonomic Dysfunction
	SFT	Semantic Fluency Test
	REM-RBD	Rapid Eye Movement Sleep Behavior Disorder
	Epworth	Epworth Sleepiness Scale
	HVLT	Hopkins Verbal Learning Test
	LNS	Letter Number Sequencing Test
	SDM	Symbol Digit Modalities Test
	Benton	Benton Judgement of Line Orientation

Table 1: List of Clinical Assessments

From PPMI's enormous collection of clinical assessment data, based on the literature study list of most contributing baseline assessments adopted by a medical professional for diagnosis of PD progression are described in Table 1. These selected lists of 15 clinical assessments capture the premature motor and non-motor symptoms experienced by PD patients to scale the severity

[1] http://www.ppmi-info.org

of the disease. In this research study important covariates are extracted from these set of assessments.

3.2 Understanding of Data

PPMI monitors symptomatic progression in PD patients at regular intervals. Participation of the patient at each interval is marked by a unique visit id. During every follow-up visit, patient is re-accessed for all baseline clinical assessments to capture the change in posed symptoms and understand the scenario of progression for that individual patient. These scheduled visits are marked as BL (Basel Line) (Prashanth and Dutta Roy, 2018), which is first-time evaluation of the patient, later to which each incremental visit is marked between V01 – V12.

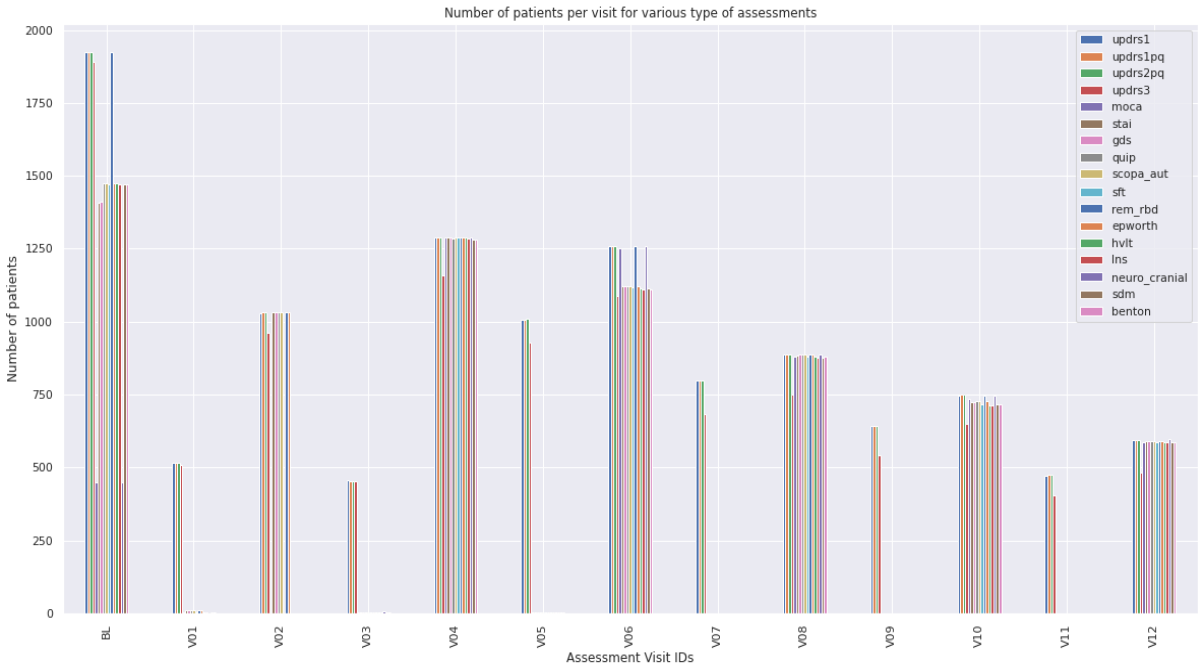

Figure 3: Participation of patients per visit for each clinical assessment

Bar plot represented in Figure 3 quantifies the number of participations per visits for a selected set of clinal assessments. Based on this data visualisation it can be inferred that participation of patients declines gradually moving towards the last visit. Also, not each assessment is mandatorily conducted on every scheduled visit. For the purpose of study, subgroup of 476 patients is created who showed active participation till last visit (V12), eliminating rest of the patients who quitted the programme halfway.

Figure 4: Distribution of Patient

Bar plot in Figure 4 represents the distribution of selected subset of patients into various categories defined by PPMI. For the purpose of this study, we focus only on Healthy Control (HC) and Parkinson's Disease (PD) patients. Also, data available for rest of the category is not sufficient for analytical study.

3.3 Design Specification

A baseline evaluation for classifying risk of disease progression is inspired by various related work carried out in space of early diagnosis of PD with applied machine learning techniques. The discussion made in related works highlights the impact factor of various assessments contributing towards effective diagnosis at prodromal phase (Prashanth and Dutta Roy, 2018) and its application for training classification model to predict the PD. Choice of machine learning algorithms and training covariates for concluded process flow of project implementation is a novel approach designed after a comparative study of various Related Work. Figure 5 illustrates the graphical abstract of the process flow and the steps involved in the implementation of this project.

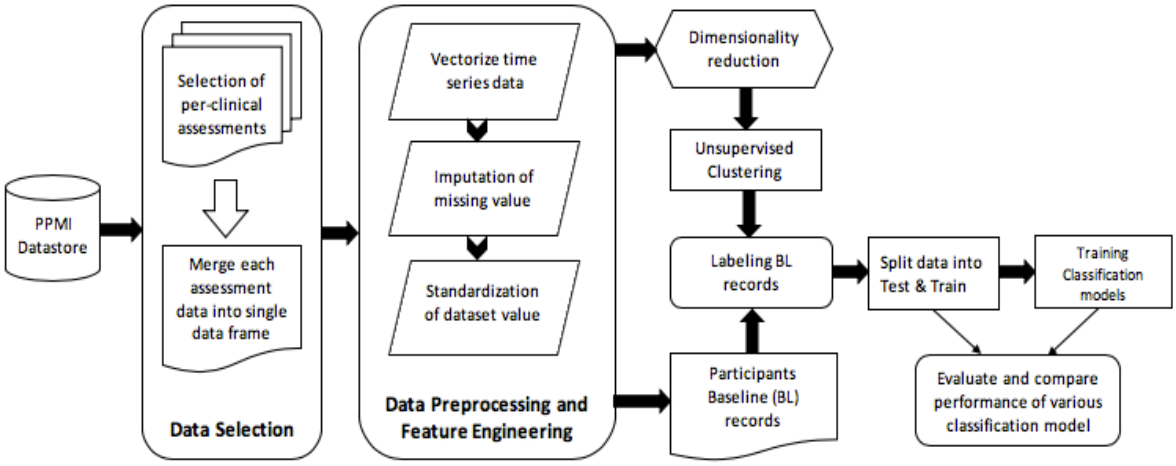

Figure 5: Implantation design flowchart

A brief description of steps involved in the designed process flow of are listed below:

1. Required clinical assessment datasets are fetched via API call from PPMI repository. Using dedicated packaged library *pypmi*[2] developed to access PPMI datastore.

2. Essential covariates are identified from the selected list of pre-clinical assessment mentioned in Table 1. Data points of these covariates are merged into a single data frame.

3. Data Pre-processing and feature engineering techniques are performed on the consolidated dataset. Time series component (visit ID's) of the dataset is vectorised into a single series. And missing values in the dataset is handles using a linear interpolation technique. Finally, cleaned data is normalized using the Min-Max scaling technique to uniformly scale data points under each covariate.

4. To scale down the number covariate, dimensionality reduction technique is applied to group correlated features into new component vectors.

5. Clustering of participants into three groups is performed based on the risk of PD progression, using an unsupervised clustering algorithm.

6. Based on these defined clusters, baseline features of patient's are labelled. This labelled dataset is split into train and test data for training and evaluation of classification models.

7. Various classification models and multi perceptron-based learning model are trained and tested with multiple set of hyper-parameters to improve the performance accuracy of the model.

8. Performance evaluation of each model against various evaluation metric and a comparative study of each model performance is conducted to identify the best performing model.

Details on the implementation of the above steps are explained in the Implementation and Evaluation section of this report. In step 5 and steps 6 dimensionally reduced feature vectors are applied to unsupervised clustering Gaussian Mixture Model (GMM). Clustering based on GMM techniques generates data clusters by grouping similar data points based on their feature and correlation between the points (Ni *et al.*, 2020). In proposed design data points are grouped into 3 clusters (Liu *et al.*, 2011), where each cluster defines the risk of PD progression among the patient. These clusters are marked as Low, Medium and High referencing to progression rate. Below is the graphical representation of the clustering component from the designed process flow in Figure 5.

[2] https://pypmi.readthedocs.io/en/latest/

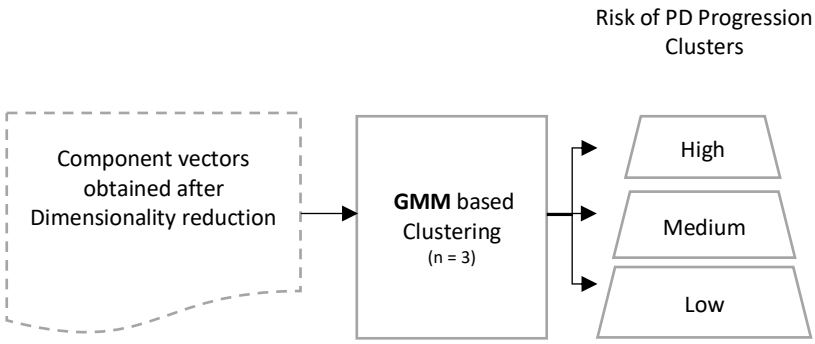

Figure 6: Process of clustering PD cases based on the risk of progression

In steps, 7 choices of classification algorithm are made based literature study of Related Work. Such as Support Vector Machine (SVM), AdaBoost, Random Forest Classifier, Decision Tree and XGBoost. In addition to that performance comparison of Multi-perceptron-based classification model is conducted against the performance of traditional classification algorithms. This study aims at evaluating the impact of a joint analysis of the selected list of clinical assessment in Table 1 for prediction of risk associated with the progression of PD using a machine learning approach. Process flow in Figure 5, is a novel designed to build a joint learning model for classification of PD patients based on their baseline features.

4 Implementation

This section elaborates the steps involved in end to end implementation of this project. Explaining the techniques adopted for data preparation and transformation, feature extraction and modelling of classification algorithms.

4.1 Environmental Setup

Implementation of this project is carried out on Google Collaboratory (Colab)[3]. It is an open-source online platform build on top of Jupiter Notebook, which allows users to run python programs on Google servers and leverages high-end GPU's and TPU's free of cost to implement machine learning model. Skipping all the hefty setup and installation for project execution, shared code file can directly be uploaded on Google Colab platform to run end to end implementation of the project.

[3] https://colab.research.google.com/notebooks/intro.ipynb

4.2 Data Preparation

Data preparation is a crucial step towards building an effective machine learning model. Selection of essential covariates from each clinical assessment was done in reference with the Data Dictionary maintained by PPMI, which describes the importance of each covariate defined in each clinical assessment dataset. Each clinical assessment dataset holds multiple records against a single patient ID equivalent to the total number of visits patient participated for assessment. Thus, each record is unique against combine index on Patient Id (*PATNO*) and visit Id (*EVENT_ID*) columns in the dataset. From Figure 3, the number of records available for each scheduled visit is highly inconsistent due to early withdrawal of participants from this study. Inconsistency and noise in data can lead to underperforming model, negatively impacting its accuracy. To eliminate this inconstancy, filtering of patient's clinical records for most promising visits ID's (*BL, V02, V04, V06, V08, V10, V12*), as identified from Figure 3 is carried out for each clinical assessment. Later to this patient ID's who quitted halfway were eliminated to filter patient ID's who participated till the last visit in all selected list of clinical assessments. This resulted in a total 476 patient ID's to progress with for building a PD classification model. Steps followed to process records of a filtered list of patients are discussed below:

- **Merge Datasets:** Records of data for all the essential covariates identified in the selected list of clinical assessment is merged over unique index key defined on patient id and visit id for the filtered list of patients.

- **Vectorise Time Series Data:** Consolidated dataset contains multiple record against individual patient corresponding every scheduled visit. To group all visits of patients into single row, the technique of vectorisation of this time series data into a single series is performed. Vectorization technique eliminates looping hops, generating better performing model and reduces the computation load by 20 – 30%. In this technique data points of each covariate from every visit of the patient is pivoted against its unique patient ID, creating a single row of record for each patient ID.

- **Data Imputation:** Data imputation is a process of handling missing values in the dataset. To handle missing values in the constructed dataset, Linear interpolation technique is adopted. Interpolation is a technique of deriving a function using discrete data points such that line of function passes through all the chosen data points. Based on this derived function value of a new point on the line or curve can be estimated, lying between two given points on the curve.

- **Data Normalization:** In a dataset it is not necessary that data points of each covariates are calculated on a same unit scale which impacts the performance of the fitted model as each variable would not contribute equally, creating training biases during model training. To handle this and bring down the data point of all covariates to a uniform scale, Min-Max normalization technique is adopted to normalize data points on a scale of 0 to 1.

- **Dimensionality Reduction:** Consolidated dataset after vectorisation of independent covariates on scheduled visit ID's have resulted in total 1595 columns. To dimensionally reduce the number of independent covariates dimensionality reduction technique is adopted. In this technique highly correlated features are grouped to form a single independent component vector. For this study, two-dimensionality reduction technique, Principle Component Analysis (PCA) and Nonnegative Matrix Factorization (NMF) are adopted to group all covariates into 2 independent component vectors. PCA extracts the linear combination of different variables that correlate in the direction with maximum sample variance. Succeeding PCA vector scales out to find a direction that gives the highest variance such that it is uncorrelated to preceding ones. On another hand, NMF performs decomposition of feature variables with an assumption that data point and the components are non-negative.

4.3 Unsupervised Clustering and Labelling

In this phase of implementation, participants are grouped into 4 defined clusters marked as Healthy Control (HC), Low is the risk of PD progression (PD_l), Medium risk of PD progression (PD_m) and High risk of PD progression (PD_h). To achieve this, unsupervised clustering is performed using a Gaussian Mixture Model (GMM) to group characteristically similar data points into the same cluster (Ni *et al.*, 2020). GMM is a probabilistic clustering model, which function on an assumption that there is a certain number of gaussian distribution exists within correlated data points and each of this distribution represents a separate cluster.

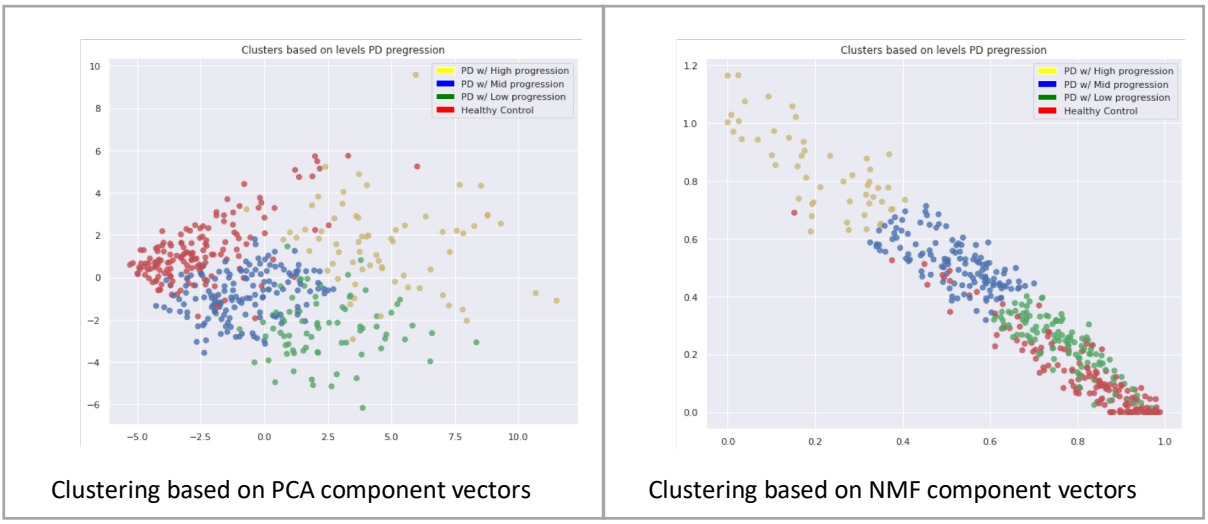

Figure 7: Participants clusters based on the risk of progression

Clustering model is fitted separately on component vectors generated by PCA and NMF dimensionality reduction technique. Figure 7 illustrates the comparative representation of resultant clusters generated by GMM model in both cases. It can be inferred that in case of clustering over PCA components, both positively as well as negatively correlated components

are grouped together whereas in case of NMF tends to find patterns among variable with the same direction of correlation.

To acknowledge the objective raised in proposed research question, baseline (BL) characteristic of participants is extracted out from min-max normalized dataset of all scheduled visits into a separate data frame. Each patient ID (PATNO) is labelled according to the patients outlined cluster. As illustrated in implementation design process flow chart, Figure 5. This labelled BL characteristic data of each participant is used to train classification models.

4.4 Modelling

This section discusses the steps for modelling and tuning of various classification model. In this study, 7 different classification models are trained. Random Forest classifier, Decision Tree, Support Vector Machine, AdaBoost, XGBoost and Multi-layer Perceptron classification model. Performance of each model is evaluated to identify the best fitting model for classification of the proposed problem statement. Participants labelled BL characteristic data is spitted into train and test set at a split ratio of 0.2, leaving 358 records in the training set and 90 records in the test set.

Multilayer-Perceptron learning model:

Based on the literature study in Related Work, architecture for Multilayer perceptron (MLP) learning model is built. Choice of hyper-parameters for the designed model is a state of art achieved after testing multiple sets of hyper-parameter combinations. In the finalized model, input dimension of BL features is 235. The model contains 2 Dense hidden layers with ReLu activation function. Dropout is used to randomly turn off neurons between layers to improve performance regularization of the model and avoid overfitting. Dropout ratio of 0.5 is used between two hidden layers. Output dimension consists of 4 nodes representing 4 classes of the classification model. The SoftMax function is used in the output layer to receive probability distribution-based prediction across each class. Figure 8 illustrates the flow diagram of neural network based multi-class classification model.

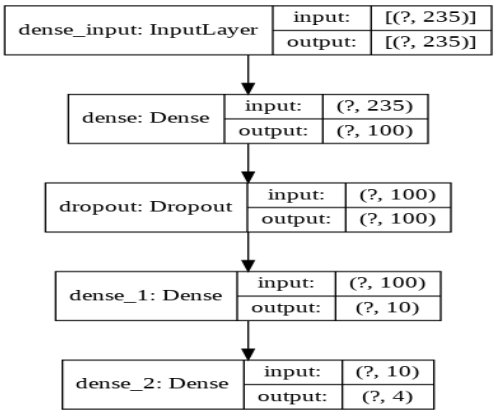

Figure 8: Multilayer Perceptron model architecture

Built model is compiled using '*adam'* optimizer and '*sparse_categorical_crossentropy*' loss function. For training and validation of the MLP model, 25 per cent of training data is spitted used as validation data. Training of the MLP model is performed with a batch size of 32 over 50 epochs. However, Early stopping function is applied with the patience of 5 epochs to stop training model in case of overfitting. Also, Model checkpoint function to save the best performing model. These training functions helps to monitor the training rate and prevent overfitting in the built model caused due to excessive epochs.

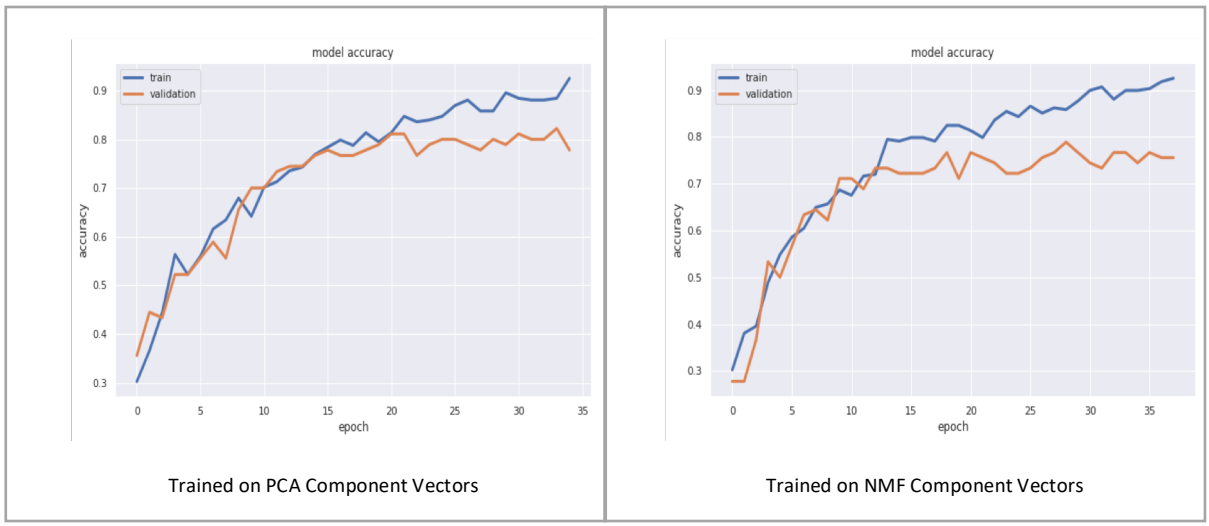

Figure 9: Training accuracy of Multi-Layer Perceptron model

Neural network models take up incremental learning approach. Learning curves are widely used to evaluate training and validation performance of the model. Accuracy curve and losses curve helps to diagnose underfitting and overfitting in the trained model. Based on which hyperparameter tuning techniques can be adopted to improve model performance. Performance of MLP model doesn't deviate much between PCA and NMF component vector. Figure 9 illustrates model training and validation accuracy curve. In both the cases, it can be depicted that training and validation accuracy is increasing gradually per epoch and reaches a plateau. Overall, good training accuracy is achieved in both scenarios. Based on these both curve formation we can state that model is not overfitted nor under fitted.

Component Vectors / Metrics	**Accuracy**	**MCC**
NMF	84%	0.77
PCA	81%	0.73

Table 2: MLP model performance comparison

Evaluation of MLP model on test data result in minor difference between models trained over NMF and PCA component vector. From table 2, Model trained on NMF component vector outperforms PCA based model, resulting higher classification accuracy of 84% and Mathews Correlation Coefficient score of 0.77.

5 Evaluation

This section explains the various evaluation metrics employed to evaluate the classification models. Also, the performance of these classification models is compared for two case scenarios in respect with employed dimensionality reduction technique, PCA and NMF. Each model is trained separately with clustering obtained from PCA component vectors and NMF component vectors. Based on continuous evolution and tuning the models, prediction accuracy for best fitting models are listed in Table 3.

Model Classification Accuracy		
Models	**PCA**	**NMF**
SVM	77%	86.7%
MLP	81%	84%
XGBoost	82%	80%
AdaBoost	70%	61%
Random Forest	81%	77%
Decision Tree	77%	68%
Nearest Neighbors	60%	53%

Table 3: Performance comparison of classification models

From Table 3, It can be inferred that the performance of classification models varies widely when trained over PCA and NMF based clustering data. Also, from the described accuracy score for each model, it can be inferred that SVM model delivers highest classification accuracy in comparison to another ensemble models, with an accuracy score of 87% for NMF based clustering scenario. However, it does not perform well for PCA based clustering scenario. XGBoost ensemble learning model delivers highest accuracy for PCA based clustering scenario, with 82% prediction accuracy. Down the line MLP model deliverers good results in both scenarios with performance accuracy of 81% and 84%. Out of all Nearest Neighbours algorithm comes out to be the poorest performing model for classification on fitted data points in both scenarios with an accuracy score of 60% and 53%.

Model	Metrics	MCC	Test Accuracy	F1-Score	Recall	Precision
SVM	**NMF**	0.82	0.87	0.86	0.87	0.87
XGBoost	**PCA**	0.75	0.82	0.81	0.82	0.82

Table 4: Evaluation scores for best performing models

Further evaluation of the best preforming MLP model is conducted. Table 4 describes the evaluation score various metrics calculated for the SVM and XGBoost model. Mathews Correlation Coefficient (MCC) score is the measure of correlation between observed and predicted classification values. MCC score close to 1 resembles a good classification model

and negative correlation values or 0 resembles a poorly fitted model with random predictions (Saqlain *et al.*, 2019). MCC is evaluated from the confusion matrix. For SVM model trained over NMF based clustering scenario delivers better MCC score of 0.82 when compared to XGBoost model from PCA based clustering scenario. Also, values of evaluation metric Precision and Recall is higher than for SVM model. The precision score gives the ratio of in respect to the total number of positives in the result. Whereas, Recall score measures the proportion of actual positives that are rightly predicted. Another evaluation metric widely used is F1-Score, It is the harmonic mean calculated from precision and recall score. It ranges between 0 -1, higher the values of F1-score better is the classification performance (Saqlain *et al.*, 2019) of the model. Which in our case is 0.86 for the best performing SVM model.

6 Discussion

This research gives insight on the application of GMM based clustering of PD patients into 3 clusters over PCA and NMF dimensionality reduced vector component. For purpose of study, comparative evaluation of Multilayer Perceptron learning model with six other traditional classification models, SVM, AdaBoost, XGBoost, Random Forest and Decision Tree for the selection the best performing prediction model. After end-to-end evaluation of each model, based on results obtained in Table 3, it can be concluded that SVM model delivers best classification accuracy of 87% when clustering is performed over NMF based dimensionality reduced vector components. Also, it can be concluded that NMF based clustering delivers better performance accuracy for models compared to PCA based clustering. Further evaluation results of the SVM model in Table 4, the MCC score recorded for this model of 0.82 states that model has high prediction accuracy in each class.

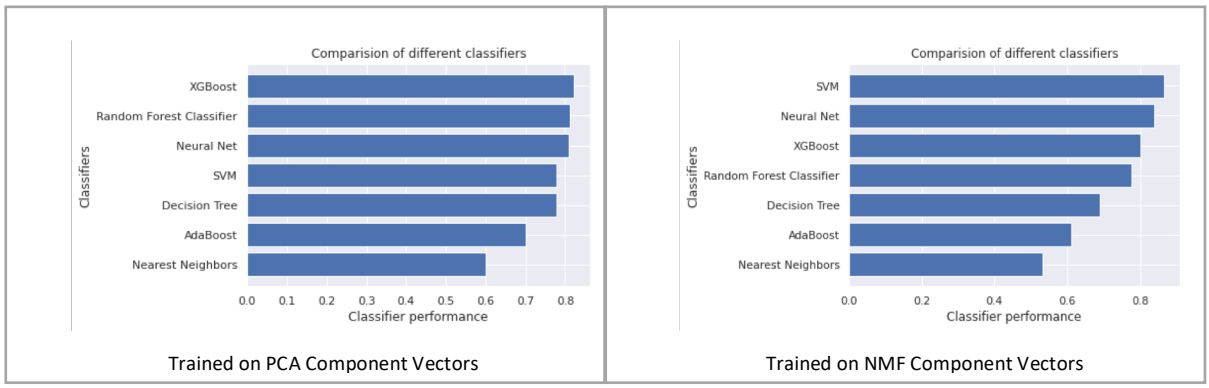

Figure 10: Process of clustering PD cases based on the risk of progression

Horizontal bar plot in Figure 10 gives the visual representation for comparison of the classification accuracy score calculated on test data for each model trained over NMF and PCA based clustering data points. Based on conducted research study it can be concluded that joint analysis of biomarkers from multiple clinical assessments can help diagnose the rate of PD progression with 87% accuracy using SVM learning technique.

7 Conclusion and Future Work

In this project, research is conducted to classify the rate of progression based on baseline evaluation of PD biomarkers responsible for cognitive and motor impairment in patients at the prodromal stage. A detailed literature study is carried out to identify the list of most suitable clinical assessments to cluster out the different categories of patients using GMM unsupervised clustering algorithms. Patients were categorised into three clusters based on the rate of progression (low, medium, high). Labelling patient based on these clusters supervised machine learning model are trained. Over performance comparison of various classification algorithms, the SVM is identified to be the best performing model with an accuracy of 87% for prediction of new patient's category based on baseline line feature evaluation.

In respect to the variety of data available by PPMI, conducted research considers a subset from enormous data, due to lack of computational resources and storage capacity. In future work. it will be interesting to see how joint learning with gene sequence data of patients can impact the decision-making accuracy of this classification model. Also, various other categories of the patient which were excluded from the current study due to insufficient data can be included with a larger available dataset. Finally, taking into consideration the available resource and implementation constraints for the conducted research, experimental results and findings from this research will contribute to future research work in area PD progression.

Acknowledgement

I would like to take this opportunity to pay my gratitude to everyone who has supported me in completing this research project successfully on time and as planned. I would like to convey my sincere gratitude to my supervisor Prof. Christian Horn for his constant support and keeping me motivated throughout research project implementation. Under his supervision and feedback, I was able to carry out my research smoothly. Also, I would like to thank my fellow batchmates under his supervision for engaging in healthy discussions and suggestion regarding the implementation and reporting of conducted research in the past few weeks. I would also like to thank National college of Ireland for providing support and assistance with NCI Library providing enormous sources for literature and guidance in report writing. Lastly, I would like to thank my family and friends for their support and believing in me.

References

Bernardo, L. S. *et al.* (2019) 'Handwritten pattern recognition for early Parkinson's disease diagnosis', *Pattern Recognition Letters*, 125, pp. 78–84. doi: 10.1016/j.patrec.2019.04.003.

Bhat, S. *et al.* (2018) 'Parkinson's disease: Cause factors, measurable indicators, and early diagnosis', *Computers in Biology and Medicine*, pp. 234–241. doi: 10.1016/j.compbiomed.2018.09.008.

Celik, E. and Omurca, S. I. (2019) 'Improving Parkinson's disease diagnosis with machine learning methods', in *2019 Scientific Meeting on Electrical-Electronics and Biomedical Engineering and Computer Science, EBBT 2019*. doi: 10.1109/EBBT.2019.8742057.

Challa, K. N. R. *et al.* (2017) 'An improved approach for prediction of Parkinson's disease using machine learning techniques', in *International Conference on Signal Processing, Communication, Power and Embedded System, SCOPES 2016 - Proceedings*, pp. 1446–1451. doi: 10.1109/SCOPES.2016.7955679.

Driver-Dunckley, E. *et al.* (2014) 'Olfactory dysfunction in incidental Lewy body disease and Parkinson's disease', *Parkinsonism and Related Disorders*, 20(11), pp. 1260–1262. doi: 10.1016/j.parkreldis.2014.08.006.

Faivre, F. *et al.* (2019) 'The hidden side of Parkinson's disease: Studying pain, anxiety and depression in animal models', *Neuroscience and Biobehavioral Reviews*, pp. 335–352. doi: 10.1016/j.neubiorev.2018.10.004.

Goetz, C. G. *et al.* (2008) 'Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results.', *Movement disorders : official journal of the Movement Disorder Society*, 23(15), pp. 2129–70. doi: 10.1002/mds.22340.

Hindle, J. V. (2010) 'Ageing, neurodegeneration and Parkinson's disease', *Age and Ageing*, pp. 156–161. doi: 10.1093/ageing/afp223.

Iakovakis, D. *et al.* (2019) 'Early Parkinson's Disease Detection via Touchscreen Typing Analysis using Convolutional Neural Networks', in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 3535–3538. doi: 10.1109/EMBC.2019.8857211.

Kuslansky, G. *et al.* (2004) 'Detecting dementia with the Hopkins Verbal Learning Test and the Mini-Mental State Examination', *Archives of Clinical Neuropsychology*, 19(1), pp. 89–104. doi: 10.1016/S0887-6177(02)00217-2.

Liu, P. *et al.* (2011) 'Clinical heterogeneity in patients with early-stage Parkinson's disease: A cluster analysis', *Journal of Zhejiang University: Science B*, 12(9), pp. 694–703. doi: 10.1631/jzus.B1100069.

Meara, J., Mitchelmore, E. and Hobson, P. (1999) 'Use of the GDS-15 geriatric depression scale as a screening instrument for depressive symptomatology in patients with Parkinson's disease and their carers in the community', *Age and Ageing*, 28(1), pp. 35–38. doi: 10.1093/ageing/28.1.35.

Nasreddine, Z. S. *et al.* (2005) 'The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment', *Journal of the American Geriatrics Society*, 53(4), pp. 695–699. doi: 10.1111/j.1532-5415.2005.53221.x.

Ni, L. *et al.* (2020) 'Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model', *Journal of Hydrology*, 586. doi: 10.1016/j.jhydrol.2020.124901.

Pedrosa, T. Í. *et al.* (2018) 'Machine learning application to quantify the tremor level for Parkinson's disease patients', in *Procedia Computer Science*, pp. 215–220. doi: 10.1016/j.procs.2018.10.031.

Poewe, W. *et al.* (2017) 'Parkinson disease', *Nature Reviews Disease Primers*, 3, pp. 1–21. doi: 10.1038/nrdp.2017.13.

Prashanth, R. and Dutta Roy, S. (2018) 'Novel and improved stage estimation in Parkinson's disease using clinical scales and machine learning', *Neurocomputing*, 305, pp. 78–103. doi: 10.1016/j.neucom.2018.04.049.

Salmanpour, M. R. *et al.* (2019) 'Optimized machine learning methods for prediction of cognitive outcome in Parkinson's disease', *Computers in Biology and Medicine*, 111. doi: 10.1016/j.compbiomed.2019.103347.

Saqlain, S. M. *et al.* (2019) 'Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines', *Knowledge and Information Systems*, 58(1), pp. 139–167. doi: 10.1007/s10115-018-1185-y.

Singhal, A. *et al.* (2020) 'Modeling and prediction of COVID-19 pandemic using Gaussian mixture model', *Chaos, Solitons and Fractals*, 138. doi: 10.1016/j.chaos.2020.110023.

Soltaninejad, S., Basu, A. and Cheng, I. (2019) 'Automatic Classification and Monitoring of Denovo Parkinson's Disease by Learning Demographic and Clinical Features', in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 3968–3971. doi: 10.1109/EMBC.2019.8857729.

Tsiouris, K. M. *et al.* (2017) 'Predicting rapid progression of Parkinson's Disease at baseline patients evaluation', in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 3898–3901. doi: 10.1109/EMBC.2017.8037708.

Tsiouris, K. M. *et al.* (2018) 'A decision support system based on rapid progression rules to enhance baseline evaluation of Parkinson's disease patients', in *2018 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2018*, pp. 329–332. doi: 10.1109/BHI.2018.8333435.

Xiong, Y. and Lu, Y. (2020) 'Deep Feature Extraction from the Vocal Vectors Using Sparse Autoencoders for Parkinson's Classification', *IEEE Access*, 8, pp. 27821–27830. doi: 10.1109/ACCESS.2020.2968177.

Yang, H. J. *et al.* (2019) 'Measuring anxiety in patients with early-stage Parkinson's disease: Rasch analysis of the state-trait anxiety inventory', *Frontiers in Neurology*, 10(FEB). doi: 10.3389/fneur.2019.00049.