Applied Computer Science, vol. 20, no. 3, pp. 171–191
doi: 10.35784/acs-2024-35

Submitted: 2024-08-21 | Revised: 2024-09-08 | Accepted: 2024-09-24

Keywords: Ensemble Learning, Machine Learning, neurodegeneration, Parkinson's disease

Anitha Rani PALAKAYALA [0000-0001-7020-0284]*, *Kuppusamy P* [0000-0001-5369-8121]*

A QUALITATIVE AND QUANTITATIVE APPROACH USING MACHINE LEARNING AND NON-MOTOR SYMPTOMS FOR PARKINSON'S DISEASE CLASSIFICATION: A HIERARCHICAL STUDY

Abstract

Parkinson's Disease (PD) is a neurodegenerative disorder that impacts movement, speech, dexterity, and cognition. Clinical assessments primarily diagnose PD, but symptoms' variability often leads to misdiagnosis. This study examines ML algorithms to distinguish Healthy People (HP) from People with Parkinson's Disease (PPD). Data from 106 HP and 106 PPD participants, who underwent the Parkinson's Disease Sleep Test (PDST), Hopkin's Verbal Learning Test (HVLT), and Clock Drawing Test (CDT) from the Parkinson's Progression Markers Initiative (PPMI) were used. A custom HYBRID dataset was also created by integrating these 3 datasets. Various Machine Learning (ML) Classification Algorithms (CA) were also studied: Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), and Logistic Regression (LR). Multiple feature sets: the first quartile (Q1: 25 % most important features), second quartile (Q2: 50 % most important features), third quartile (Q3: 75 % most important features), and fourth quartile (Q4: All 100 % features) were generated using various Feature Selection (FS) algorithms and ensemble mechanisms. Results showed that all the ML CA achieved over 73±8.4 % accuracy with individual datasets, while the proposed HYBRID dataset achieved a remarkable accuracy of 98±0.6 %. This study identified the optimal quantity of non-motor features, dataset, the best FS and CA in hierarchical approach for early PD diagnosis and also proved that PD may be diagnosed with great accuracy by analyzing non-motor PD parameters using ML algorithms. This suggests that extended data collection could serve as a digital biomarker for PD diagnosis in the future.

1. INTRODUCTION

Parkinson's disease is the second most prevalent neurodegenerative disorder after Alzheimer's disease and over 6 million people worldwide got affected with PD (De Lau & Breteler, 2006). The symptoms of Parkinson's disease gradually worsen, causing a significant decrease in quality of life (Schrag et al., 2000) and life expectancy. Although there is no cure, pharmacological and surgical treatments can effectively manage PD

* VIT-AP University, School of Computer Science and Engineering, India, anitha.palakayala@gmail.com, drpkscse@gmail.com

symptoms (Connolly & Lang, 2014). Diagnosing PD is challenging due to its symptom overlap with other movement disorders and the fluctuating nature of PD symptoms (Pahwa & Lyon, 2010). ML approaches have been steadily used over the past few years for early diagnosis of PD. This has improved prediction accuracy significantly employing a variety of data types. These modalities include handwritten patterns (Drotár et al., 2014), audio signals (Sakar et al., 2013), neuroimaging methods (Nuvoli et al., 2020), and biofluids (Adeli et al., 2016). There is still a lack of studies utilizing non-motor symptoms for PD diagnosis using ML approaches. From the recent works applying ML it has been observed that out of 211 publications, 170 focused on distinguishing PPD from HP. Prashanth et al. (2014) and Mabrouk et al. (2018) used exclusively non-motor symptoms of PD resulting in accuracy of 85.48 %, 82.2 % respectively. The study proposed by Armañanzas et al. (2013) aimed to assess PD severity rather than diagnosis. The use of ML algorithms based on non-motor symptoms in clinical settings may be encouraged to support general practitioners in making decisions as their influence on the healthcare system will be enhanced by designing the models that are understandable and clear (Vellido, 2020).

1.1. Contributions

The major contributions made by this research for improving prediction accuracy are as follows:

- *Development of Biomarker for PD Diagnosis*: The authors developed a model that combines qualitative and quantitative feature selection methods with ML, enhanced the accuracy of PD diagnosis and also validated with the Bayesian Correlated t-test (BCT), thus devising a biomarker for PD diagnosis.
- *Hierarchical approach*: The authors designed a hierarchical approach that identifies the best feature selection algorithm generating qualitative non-motor features in the first level, the best ML model in the next level that could better differentiate PPD from HP using the result of level-1 and finally identifying the best dataset in level-3.
- *Evidence Aggregation Model (EAM)*: This novel Evidence Aggregation Model integrates diverse symptom assessments from multiple test types into a unified diagnostic score. This approach improves differentiating PPD from HP.
- *Enhanced Predictive Efficacy*: The proposed methodology achieved substantial improvements in predictive accuracy compared to several robust baseline models. These findings are crucial for clinical practice, offering clinicians more precise diagnostic tools for PD diagnosis.

These contributions highlight the efficacy of this approach in leveraging ML and FS techniques to advance the diagnosis and management of PD.

2. RELATED WORKS

Govindu and Palwe (2023) explored various ML models including K-nearest neighbours (KNN), logistic regression, random forest regression, and SVM for classifying PD using audio data. Their findings indicated that the KNN model achieved the highest accuracy of 91.83 %. Moradi et al. (2022) used SVM models to predict the start of PD in elderly people, with a baseline accuracy of 88.9 % using genetic data. The revised SVM model presented in

this work highlights the superiority of audio data over genetic data in the classification of PD, with an accuracy of 91.83 %. Raundale et al. (2021) utilized keystroke data to predict PD severity in older patients through a RF classifier. Cordella et al. (2021) focused on audio data to classify Parkinsonian tremor, utilizing MATLAB extensively for model development. Ali et al. (2022) applied ensemble techniques on voice data. Their study highlighted the importance of feature selection by implementing PCA to enhance the model's performance in detecting major speech differences relevant to PD. A decision tree was trained using 12 unique vocal characteristics from the dataset by Huang et al. (2021) in an effort to reduce the dependency on wearable technology for PD diagnosis. Wodzinski et al. (2019) used a Deep Learning (DL) model on audio data images for PD classification, focusing on capturing frequency nuances. Wroge et al. (2018) developed an unbiased ML model achieving a peak accuracy of 85% for predicting PD, aiming to minimize subjective diagnosis by doctors. Wang et al. (2020) implemented various ML models on a speech dataset, achieving 96.45% accuracy with a custom DL model, albeit with high memory requirements. Alkhatib et al. (2020) achieved 95% accuracy using linear classification to characterize PD patient gait, suggesting integration of audio and sleep data for improved diagnostics. Ricciardi et al. (2020) used decision trees, random forest, and KNN on brain MRI scans for Mild Cognitive Impairment detection in PD patients, utilizing artificial data augmentation due to dataset limitations. Haq et al. (2019) applied L1-support SVM on vowel phonation data from neurological disorder patients aged 46-85 years, focusing on classification without explicit feature identification. Mei et al. (2021) emphasized ML's role in detecting PD, particularly in capturing subtle non-motor symptoms often overlooked in subjective evaluations by doctors, based on a comprehensive review of 209 studies. Another study by Smyth et al. (2023) achieved high specificity in PD detection but faced limitations in sample size and scope regarding subcortical data and long-term outcomes. Martinez-Eguiluz et al. (2023) analyzed nine ML models on non-motor features from PPMI and Biocruces databases, highlighting SVM and Multi-Layer Perceptron models with best performance of 87.5% and 86.9% accuracy, respectively. They advocated for combining datasets to improve performance but noted challenges in unified analysis due to dataset heterogeneity. The presented study addresses these gaps by aggregating data of multiple tests conducted on a group of people into a single dataset, aiming to differentiate PPD from HP, with enhanced diagnostic accuracy over single test dataset. The following Section describes the proposed model of the study, various types of datasets and their characteristics, the FS mechanisms and CA along with the evaluation process. Section 4 presents the results, discussions and comparisons of the proposed study. The paper concludes with final remarks and limitations in Section 5.

3. PROPOSED SYSTEM

In this study, a variety of clinical scales measuring non-motor characteristics, such as visuospatial ability, sleep activity, and memory impairment, were chosen. The authors aggregated data from three tests, PDST, HVLT, and CDT conducted on 106 PPD and 106 HP participants to create a custom dataset called the HYBRID dataset.

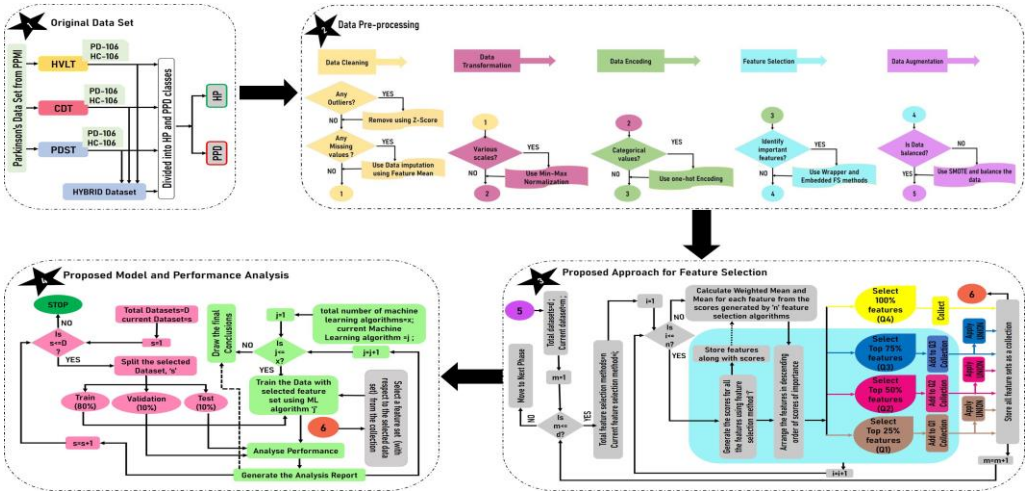

Fig. 1. The complete architecture of the proposed model

They analyzed whether this data aggregation could be beneficial to their goals. They evaluated the performance of four distinct CA. The proposed analysis was conducted in three levels with increasing granularity. Level 1 identifies the best feature selection mechanism from Univariate Selection (US), Recursive Feature Elimination (RFE), Recursive Feature Addition (RFA), Random Forest (RF) and the ensemble algorithms. Level 2 determines the best CA from RF, NB, SVM, and LR in combination with the FS algorithm selected in Level 1. Finally, Level 3 identifies the best non-motor qualitative and quantitative feature dataset for achieving a high prediction rate. Figure1 represents the overall architecture of the proposed model. Phase 1 shows the data sets used in the proposed model while phase 2 is involved with the data preprocessing mechanisms employed on the collected data. Phase 3 totally involves with the generation of different quantitative feature sets to be used in the next phase. For this purpose, four different FS mechanisms namely US, RFA, RFE and RF were used to generate scores for all the features. These are sorted in decreasing order of importance and the quartile feature sets: Q1, Q2, Q3 and Q4 respectively are generated. UNION of sets is applied on the respective quartile sets obtained from the four FS algorithms to generate four new quantitative sets, thus giving importance to each and every feature. An ensemble mechanism is applied on the scores of the respective feature generated by the FS algorithms to calculate the Mean and Weighted Mean values. These are also sorted in decreasing order of importance and the quantitative feature sets of Q1, Q2, Q3 and Q4 respectively are generated using mean and weighted mean. This entire process can be seen from phase 3 of Fig.1. Finally, phase 4 represents various ML algorithms used for training the collected data using various quantitative feature sets obtained in phase 3. The ability to identify PD is assessed and final conclusions are drawn.

3.1. Dataset

Discovering biomarkers for PD detection and progression is the goal of the large, multi-centre, long-term PPMI investigation, aiming towards improving analytical and clinical research. People who were considered early PD patients (with duration of no more than two years) and drug- naïve at the time of enrolment were included in the study.

Tab. 1. Dataset features and description

Dataset	Feature	Meaning
CDT	Pat_ID	Unique number assigned to patient
	Age	Age of the person
	Clck2hnd	exactly 2 hands
	Clckalnu	all 1-12 present
	Clcknmrk	absence of marks
	Clcknusp	equally spaced from each other
	Clcknuin	Positioning of numbers inside
	Clcknued	numbers equally distributed
	Clckpii	one hand points to 2
HVLT	Pat_ID	Uniqu number assigned to patient
	Hvltvrsn	Version number
	Hvltrt1	Immediate Recall Trial 1
	Hvltrt2	Immediate Recall Trial 2
	Hvltrt3	Immediate Recall Trial 3
	Hvltrdly	Delayed Recall Trial 4
	Hvltrec	Total count of true positives
	Hvltfprl	Total false positives, related
	Hvltfpun	Total false positives, unrelated
	Age_Assess_Hvlt	Age at Assessment
	Dvt_Total_Recall	Derived-Total Recall T-Score
	Dvt_Delayed_Recall	Derived-Delayed Recall T-Score
	Dvt_Retention	Derived-Retention T-Score
	Dvt_Recog_Disc_Index	Derived-Recognition Discrimination Index T-Score
PDST	Pat_ID	Uniqu number assigned to patient
	Slept_Well	sleep quality
	Dfclty_Fall_Asleep	difficulties initiating sleep
	Dfclty_Stay_Asleep	staying asleep
	Pain_Post_Of_Limbs, Urge_Move_Limbs	nocturnal restless legs syndrome
	Distrssing_Dreams	vivid distressing dreaming
	Pdss_Dstrss Halluc	Hallucinations
	Pass_Urine	nocturnal urinary urgency
	Immobile	immobility at night
	Pain_In_Limbs,	Sleep related pain
	Cramps_In_Limbs	muscle cramps
	Tired_On_Wake	painful posture on wakeup
	Tremor_On_Wake	tremor on waking
	Rstlss_Leg	lack of repose from sleep
	Snoring_Woke	sleep apnea

Data from 106 PPD and 106 HP participants who have undergone PDST, HVLT, and CDT were collected from PPMI. The criteria considered is Bradykinesia being identified as the core motor feature along with at least one of the following features: stiffness, resting tremors or postural instability. All patients gave written informed consent before beginning the study while taking medication to complete all assessments, as required by the Declaration of Helsinki. Small sample size may limit the generalizability of the classification models. But our study is a small approach to know how far the individual tests of a PD patient can be used to diagnose the disease efficiently compared to the collection of tests i.e., the hybrid data. In view of this, we were able to gather only 106 PD patient's data which is available across all the three tests namely HVLT, PDST and CDT from PPMI database. This limited

our collection of Healthy people too, to 106 to maintain balance between PD and Healthy People count. We hope that this study will help the future researchers to study on a large-scale data with various other economy tests available, to diagnose PD. Hoehn and Yahr (HY) scale was used to grade the severity of the disease. The demographic and clinical characteristics of the HC and PD patients are shown in tab. 2.

Tab. 2. Characteristics of the research group

Class	Criteria	Description
PPD	Gender	Male (50) / Female (56)
	Age	Between 39 and 81 years
	Symptoms	Resting Tremor / Bradykinesia / pill rolling / Rigidity (any two)
	Diagnose duration	Two years (minimum)
	Disease stage	Hoehn & Yahr Stage I / II
HP	Gender	Male (65) / Female (41)
	Age	Between 34 and 85 years
	Symptoms	No indications of any neurological diseases

3.2. Non-Motor Test Assessments to detect PD

3.2.1. CDT

It is globally recognized as a neuropsychological screening tool with robust psychometric properties, including test-retest reliability and inter-rater reliability (Mainland & Shulman, 2017). Various scoring systems exist, with higher scores indicating better performance and lower scores suggesting potential cognitive deficits. The CDT's simplicity, quick administration, and ability to assess a range of cognitive functions have made it popular in both research and clinical settings for screening cognitive disorders. Various features of CDT are shown in Tab. 2. A sample clock drawing of 3 PPD can be seen from Fig. 2.

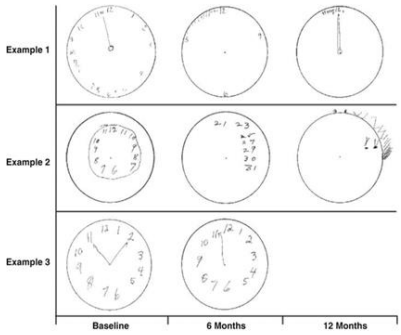

Fig. 2. Sample drawings of CDT [57]

3.2.2. HVLT

HVLT test (Benedict et al., 1998) involves three trials where participants freely recall a 12-item list categorized semantically. Following these trials, a YES/NO recognition task is conducted. Approximately after20 minutes, an impede memory test, in which participants freely recollect words from the original list and a detection test (consists of 24 words: 12 target words, 12 incorrect positives, 6 semantically similar distractors, and 6 semantically

independent distractors) are administered. The features of HVLT can be observed from Tab. 1.

3.2.3. PDST

Table 2. depicts various features associated with PDST (Thangaleela et al., 2023). PDSS-2 (Chaudhuri et al., 2002) consists of a modified version of 15 questions that assess different dimensions of sleep disturbances. Every subject is scored using a categorical scale that indicates how frequently the disturbance occurred (A score of 0 indicates not at all disturbed; 1 indicates disturbed rarely; 2 indicates disturbed few times; 3 indicates disturbed regularly and 4 indicates disturbed very frequently). The questions are related to the past week. The highest score of 60 indicates maximum disturbance while the lowest score of 0 indicates no disturbance. A score ≥18 indicates significant PD-related sleep disruptions (Trenkwalder et al., 2011).

3.2.4. HYBRID dataset

Combining heterogeneous datasets enriches the available information and enhances predictive performance by leveraging diverse data sources. This approach leads to better decision-making through a comprehensive view of the data, enables advanced analyses, and increases robustness by averaging out noise and reducing biases. It also fosters innovation and competitive advantage by uncovering new insights and making models more flexible and adaptable to various tasks. In the proposed study, the three diverse datasets CDT, PDST and HVLT conducted on same cohort of people are combined as follows: Initially, all the individual datasets are cleaned by handling the missing values, correcting the errors and removing the duplicates, if any. Schema matching is performed by identifying a common feature across all datasets for integration. In this process, patient-ID is the common feature observed to integrate all the datasets together. Then, left outer join operation was applied to integrate all the datasets into a single comprehensive set, the HYBRID dataset. This dataset is observed to check the presence of any duplicate records as a result of integration and removed, if any. All the features of CDT, HVLT and PDST are found in the HYBRID making up a total of 38 features.

3.3. Data preprocessing

The data obtained from PPMI underwent several preprocessing steps, including imputation of missing values, removal of outliers, and application of rescaling techniques where necessary.

3.3.1. Data cleaning

Outliers are data points that significantly differ from the majority of the data in a dataset. Outliers can adversely affect the model performance with increased errors that can distort data visualizations, making it difficult to interpret the underlying patterns. These outliers can be detected using visual, statistical and ML methods. Outliers are identified and eliminated using Z-score method in the proposed study. The value with Z-score above a certain threshold (±3 in this study) are considered outliers and hence eliminated.

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

Where 'x' is the data point; 'μ' is the mean of the dataset and 'σ' is the standard deviation of the dataset.

The dataset used in this study was collected from the public database, PPMI and some data points were incomplete or missing. This is often a common issue with publicly available datasets, as they may contain gaps due to data collection processes or inconsistencies in reporting. The missing data could be due to incomplete entries or technical issues during the original data collection phase. As the missing data may introduce some limitations, we took necessary precautions to minimize its impact on the results. To address this, we used data imputation to fill missing values using mean, ensuring the integrity of the analysis. Data imputation is a statistical technique used to replace missing or incomplete data within a dataset with substituted values. This is done to maintain the integrity of the data and allow for accurate analysis. It is observed that the collected datasets are with some missing values and after confirming that the data approximately follows a normal distribution using the Shapiro-Wilk test, data imputation using mean was used. Each feature with missing values is observed and mean is calculated with which the gaps are replaced.

$$Mean = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (2)$$

Where 'n' is the total number of data points and $x_i$are the individual data points.

3.3.2. Data transformation

Data normalization is a preprocessing technique that adjusts data to a common scale while preserving the relative differences in value ranges. This step is particularly crucial when features have different units or scales. Min-Max normalization is a technique used to rescale the range of features to a specific range, typically [0, 1]. This method adjusts the data so that the minimum value of each feature becomes 0 and the maximum value becomes 1. It can be calculated using the formula as follows:

$$X_{norm} = \frac{X-X_{min}}{X_{max}-X_{min}} \quad (3)$$

Where 'X' is the original value; X_{min} is the minimum value of the feature; $X_{max}$is the maximum value of the feature; $X_{norm}$is the normalized value; In the proposed study, quantitative features were normalized using min-max normalization, ensuring all variables ranged between 0 and 1.

3.3.3. Feature encoding

Feature encoding is the process of converting categorical data into a numerical format that ML algorithms can understand. In this study, categorical features were encoded using one-hot encoding. It works by creating new binary columns, each representing one of the possible categories in the original data, and assigning a 1 or 0 to indicate the presence or absence of each category. This method ensures that categorical data is treated appropriately by algorithms, avoiding the creation of any artificial ordinal relationships. However, it can

lead to increased dimensionality and sparsity in the dataset when dealing with features that have many categories.

3.3.4. Feature selection

Feature selection aims to identify the most relevant variables or biomarkers for two primary purposes. Firstly, it assists in distinguishing between PPD and HP by pinpointing significant features. Secondly, Feature selection streamlines models by simplifying the data, which enhances interpretability, reduces training times, and lowers the risk of overfitting. The proposed methods for this study are RFE, RFA, RF and US, which are chosen mainly due to their minimal computational cost and independence from specific prediction models.

3.3.5. Data augmentation

An imbalanced dataset, where one class significantly outweighs the other, can negatively impact the performance of machine learning models. This is especially problematic in binary classification, where the model might become biased towards the majority class and fail to learn the minority class patterns. The model may show high accuracy simply by predicting the majority class, even if it is not correctly classifying the minority class. This leads to misleading performance metrics. One approach followed to address data imbalance problem of the proposed study is Synthetic Minority Oversampling Technique [SMOTE] (Gunakala & Shahid, 2023) that generates synthetic samples for the minority class by creating linear combinations of existing minority samples and selecting points along these combinations. This helps balance the dataset, without simply duplicating data points that may lead to overfitting. By balancing the class distribution, SMOTE allows the model to learn patterns in both the minority and majority classes. This results in better classification performance, particularly in identifying minority class instances. The impact of data augmentation can be observed from Tab. 3.

Tab. 3. Impact of data augmentation

Dataset	Data samples		Data Growth Factor	Impact on result (accuracy improvement %)
	before augmentation	after augmentation		
CDT	485	998	2.05 times	0.82
HVLT	1084	--	--	--
PDST	355	1077	3.03 times	1.22
HYBRID	1864	3344	1.79 times	1.06

3.4. Training and fitting

The ML algorithms are programmed to learn and optimize their operations by analysing input data to make predictions within defined parameters. As new data is introduced, these algorithms improve their accuracy in prediction. In this study, 4 ML algorithms NB (Zhang, 2004), RF (Breiman et al., 2001), SVM (Pisner & Schnyer, 2020) and LR (Hosmer et al., 2013) were applied on the data obtained from PPMI database as well as the custom HYBRID data sets using various quantified features. The reasons behind the choice of ML techniques are as follows:

1. Random Forest: This is an ensemble learning technique that builds multiple decision trees and merges them to produce a more accurate and stable prediction. It is especially useful in handling datasets with many features, preventing overfitting, and offering high accuracy, making it ideal for capturing complex relationships in the data.
2. Naive Bayes: Based on Bayes' Theorem, this model is fast and efficient. It assumes feature independence and performs well when a probabilistic approach is needed. It is known for handling noisy data well.
3. Logistic Regression: This is a simple and interpretable model that outputs probabilities directly, making it great for binary classification. Its effectiveness lies in its simplicity and ability to provide insights into the impact of each feature on the predicted outcome.
4. SVM (Support Vector Machine): SVM is particularly good at separating classes with a clear margin of separation, offering a robust solution for linearly separable data and kernel-based methods for more complex patterns.

Hyperparameters are the parameters set by the user before the training of a ML model begins. To optimize model performance, Grid Search was employed to identify the best hyperparameter values. This method systematically explores a grid of hyperparameter combinations, evaluating each combination using cross-validation. Grid Search is straightforward to parallelize as each hyperparameter combination is evaluated independently, making it efficient and scalable (Yu & Zhu, 2020).

4. RESULTS

4.1. Hierarchical approach

All the tasks are carried out on a laptop equipped with an Intel ® Core (TM) i5-1135G7 microprocessor, NVidia GeForce GTX 1660 GPU, 16 gigabytes of Random-Access Memory (RAM) and a 256 GB DDR4 SSD for storage. Tab. 3. represents the Confusion Matrix (CM) obtained to describe the behaviour of the ML algorithms RF, SVM, NB and LR when applied on datasets CDT, HVLT, PDST and HYBRID datasets. Each and every cell is composed of 4 values representing True Negatives (TN), False Negatives (FN), True Positives (TP) and False Positives (FP) respectively. The highlighted values represented in HYBRID dataset using RF algorithm with all feature sets, represents the confusion matrices that contributed to the best performance. The Q2 feature set generated by RF feature selection mechanism using RF classification algorithm outperformed with 464 samples being analysed as TN with 0 samples being FN; 196 samples being classified as TP with only 9 being identified as FP. The hierarchical approach of the process analysis is explained in the following sections.

Phase-1: The best FS Algorithm analysis for PD classification

In this phase, different FS mechanisms like US, RFE, RFA and RF are applied on four different data sets HVLT, PDST, CDT and HYBRID and different quartiles Q1, Q2, Q3 and Q4 of feature sets are extracted. This resulted in generating 52 feature sets out of which 48 are generated with Q1, Q2 and Q3 features generated using 4 FS algorithms, while 4 sets are generated with Q4 (i.e., no FS) features applied on 4 different data sets. Then, we performed ensemble operations like UNION, mean and weighted mean on the corresponding quartile

features of all FS mechanisms generating 36 (Q1, Q2 and Q3 are comprised of different sets of features generated by 3 different ensemble operations) more feature sets. These aggregates to 88 different feature sets, which can be observed from Tab. 4. All these are analysed using different CA like RF, NB, SVM and LR. This step identifies the best feature set in combination with the CA that yields better accuracy.

Tab. 4. Confusion matrices of various ML classification algorithms on different datasets

Quar-Tile	Dataset	CDT				HVLT				PDST				HYBRID			
	FS/CA	SVM	RF	LR	NB	SVM	RF	LR	NB	SVM	RF	LR	NB	SVM	RF	LR	NB
Q1	US	[50 9] [13 25]	[50 9] [13 25]	[50 9] [13 25]	[55 4] [33 5]	[133 4] [78 2]	[123 14] [67 13]	[136 1] [75 5]	[128 9] [65 15]	[36 2] [19 14]	[36 2] [19 14]	[36 2] [19 14]	[36 2] [21 12]	[445 19] [87 118]	[464 0] [63 142]	[445 19] [88 117]	[418 46] [73 132]
	RFE	[50 9] [11 27]	[50 9] [12 26]	[50 9] [11 27]	[52 7] [38 0]	[125 12] [48 32]	[113 24] [32 48]	[135 2] [79 1]	[128 9] [72 8]	[33 5] [14 19]	[34 4] [14 19]	[33 5] [14 19]	[37 1] [23 10]	[443 21] [80 125]	[457 7] [21 184]	[447 17] [95 110]	[438 26] [88 117]
	RFA	[56 3] [37 1]	[56 3] [37 1]	[56 3] [37 1]	[53 6] [35 3]	[133 4] [71 9]	[123 14] [68 12]	[137 0] [80 0]	[135 2] [80 0]	[34 4] [21 12]	[28 10] [20 13]	[34 4] [21 12]	[36 2] [24 9]	[439 25] [157 48]	[449 15] [38 167]	[431 33] [153 52]	[402 62] [127 78]
	RF	[50 9] [13 25]	[50 9] [13 25]	[50 9] [13 25]	[55 4] [33 5]	[125 12] [48 32]	[115 22] [34 46]	[135 2] [79 1]	[128 9] [72 8]	[34 4] [15 18]	[31 7] [14 19]	[34 4] [15 18]	[38 0] [23 10]	[443 21] [80 125]	[456 8] [22 183]	[447 17] [95 110]	[438 26] [88 117]
	UNION	[52 7] [13 25]	[49 10] [10 28]	[52 7] [13 25]	[49 10] [32 6]	[130 7] [57 2]	[118 19] [41 39]	[132 5] [66 14]	[115 22] [56 24]	[36 2] [21 12]	[34 4] [18 15]	[36 2] [21 12]	[36 2] [19 14]	[445 19] [87 118]	[464 0] [64 141]	[439 25] [85 120]	[398 66] [69 136]
	mean	[50 9] [9 29]]	[50 9] [13 25]	[50 9] [9 29]	[49 10] [31 7]	[125 12] [63 17]	[104 33] [56 24]	[135 2] [76 4]	[127 10] [67 13]	[34 4] [15 18]	[32 6] [16 17]	[34 4] [15 18]	[37 1] [22 11]	[445 19] [82 123]	[463 1] [11 194]	[442 22] [87 118]	[396 68] [73 132]
	Wt-mean	[50 9] [9 29]]	[49 10] [12 26]	[50 9] [9 29]	[49 10] [32 6]	[131 6] [70 10]	[104 33] [61 19]	[137 0] [78 2]	[133 4] [73 7]	[31 7] [15 18]	[28 10] [14 19]	[31 7] [15 18]	[36 2] [23 10]	[449 15] [91 114]	[464 0] [63 142]	[444 20] [86 119]	[402 62] [73 132]
Q2	US	[51 8] [19 19]	[49 10] [13 25]	[51 8] [19 19]	[49 10] [32 6]	[127 10] [63 17]	[111 26] [48 32]	[128 9] [63 17]	[120 17] [58 22]	[34 4] [15 18]	[33 5] [15 18]	[34 4] [15 18]	[36 2] [24 9]	[448 16] [93 112]	[456 8] [27 178]	[445 19] [88 117]	[418 46] [73 132]
	RFE	[52 7] [15 23]	[49 10] [12 26]	[52 7] [15 23]	[53 6] [32 6]	[135 2] [50 30]	[125 12] [35 45]	[130 7] [65 15]	[119 18] [63 17]	[36 2] [21 12]	[34 4] [18 15]	[36 2] [21 12]	[36 2] [19 14]	[448 16] [87 118]	[464 0] [17 188]	[443 21] [89 116]	[416 48] [73 132]
	RFA	[59 0] [38 0]	[56 3] [36 2]	[59 0] [38 0]	[53 6] [34 4]	[129 8] [56 24]	[117 20] [46 34]	[136 1] [79 1]	[120 17] [73 7]	[32 6] [16 17]	[33 5] [14 19]	[32 6] [16 17]	[36 2] [21 12]	[447 17] [92 113]	[463 1] [15 190]	[442 22] [90 115]	[401 63] [67 138]
	RF	[52 7] [15 23]	[50 9] [12 26]	[52 7] [15 23]	[53 6] [32 6]	[135 2] [47 33]	[124 13] [38 42]	[131 6] [76 4]	[121 16] [68 12]	[29 9] [13 20]	[29 9] [13 20]	[29 9] [13 20]	[36 2] [23 10]	[448 16] [89 116]	[464 0] [9 196]	[441 23] [89 116]	[400 64] [70 135]
	UNION	[50 9] [9 29]	[50 9] [13 25]	[50 9] [9 29]	[50 9] [31 7]	[136 1] [53 27]	[126 11] [42 38]	[130 7] [65 15]	[109 28] [55 25]	[32 6] [16 17]	[33 5] [14 19]	[[32 6] [16 17]	[36 2] [21 12]	[448 16] [92 113]	[458 6] [28 177]	[443 21] [88 117]	[397 67] [68 137]
	mean	[50 9] [9 29]	[50 9] [13 25]	[50 9] [9 29]	[49 10] [31 7]	[125 12] [58 22]	[109 28] [40 40]	[126 11] [66 14]	[123 14] [58 22]	[35 3] [14 19]	[31 7] [13 20]	[35 3] [14 19]	[36 2] [19 14]	[448 16] [93 112]	[457 7] [21 184]	[446 18] [86 119]	[397 67] [68 137]
	Wt-mean	[50 9] [9 29]	[49 10] [12 26]	[50 9] [9 29]	[49 10] [32 6]	[132 5] [55 25]	[123 14] [44 36]	[131 6] [65 15]	[122 15] [58 22]	[34 4] [16 17]	[32 6] [13 20]	[34 4] [16 17]	[36 2] [19 14]	[450 14] [94 111]	[464 0] [13 192]	[441 23] [87 118]	[397 67] [68 137]
Q3	US	[50 9] [12 26]	[50 9] [15 23]	[50 9] [12 26]	[50 9] [31 7]	[130 7] [65 15]	[119 18] [49 31]	[128 9] [64 16]	[93 44] [38 42]	[34 4] [13 20]	[33 5] [13 20]	[34 4] [13 20]	[36 2] [19 14]	[447 17] [89 116]	[464 0] [21 184]	[443 21] [88 117]	[429 35] [80 125]
	RFE	[50 9] [9 29]	[51 8] [13 25]	[50 9] [9 29]	[49 10] [30 8]	[135 2] [54 26]	[123 14] [40 40]	[132 5] [65 15]	[100 37] [42 38]	[33 5] [14 19]	[32 6] [13 20]	[33 5] [14 19]	[36 2] [19 14]	[49 15] [93 112]	[464 0] [14 191]	[443 21] [92 113]	[402 62] [75 130]
	RFA	[55 4] [33 5]	[52 7] [34 4]	[55 4] [33 5]	[47 12] [30 8]	[130 7] [58 22]	[125 12] [48 32]	[134 3] [74 6]	[109 28] [58 22]	[20 18] [17 16]	[20 18] [14 19]	[20 18] [17 16]	[32 6] [24 9]	[446 18] [92 113]	[464 0] [13 192]	[443 21] [86 119]	[415 49] [75 130]
	RF	[50 9] [9 29]	[50 9] [13 25]	[50 9] [9 29]	[49 10] [31 7]	[136 1] [59 21]	[118 19] [41 39]	[128 9] [64 16]	[89 48] [40 40]	[29 9] [16 17]	[29 9] [18 15]	[29 9] [16 17]	[37 1] [23 10]	[450 14] [95 110]	[464 0] [17 188]	[443 21] [91 114]	[402 62] [70 135]
	UNION	[50 9] [12 26]	[49 10] [13 25]	[50 9] [12 26]	[48 11] [30 8]	[135 2] [57 23]	[126 11] [42 38]	[131 6] [65 15]	[78 59] [35 45]	[30 8] [14 19]	[33 5] [15 18]	[30 8] [14 19]	[36 2] [23 10]	[450 14] [91 114]	[464 0] [17 188]	[441 23] [84 121]	[410 54] [70 135]
	mean	[50 9] [9 29]	[50 9] [13 25]	[50 9] [9 29]	[49 10] [31 7]	[133 4] [64 16]	[125 12] [50 30]	[129 8] [64 16]	[94 43] [37 43]	[36 2] [19 14]	[36 2] [19 14]	[36 2] [19 14]	[36 2] [21 12]	[449 15] [93 112]	[464 0] [14 191]	[444 20] [89 116]	[398 66] [69 136]
	Wt-mean	[50 9] [9 29]	[49 10] [12 26]	[50 9] [9 29]	[49 10] [32 6]	[130 7] [56 24]	[124 13] [47 33]	[129 8] [64 16]	[110 27] [59 21]	[36 2] [21 12]	[34 4] [18 15]	[36 2] [21 12]	[36 2] [19 14]	[449 15] [95 110]	[464 0] [18 187]	[441 23] [87 118]	[397 67] [68 137]
Q4		[50 9] [12 26]	[49 10] [13 25]	[50 9] [12 26]	[48 11] [30 8]	[135 2] [60 20]	[127 10] [46 34]	[131 6] [66 14]	[80 57] [35 45]	[35 3] [16 17]	[31 7] [14 19]	[35 3] [16 17]	[36 2] [19 14]	[449 15] [95 110]	[464 0] [18 187]	[441 23] [87 118]	[397 67] [68 137]

*Yellow coloured area represents the best confusion matrix values obtained in the respective quartile using the best dataset.

Phase-2: The best CA analysis for PD classification

In this phase, the selected feature sets of the previous step are observed to identify the best performance yielding CA in the corresponding quartiles using the 4 data sets, which can be observed from Tab. 5. As part of level-2 evaluation, the best CA that contributes to highest performance in the respective quartile feature sets is observed.

Phase-3: The best dataset analysis for PD classification

Table 5. is constructed with the aim of analysing the best dataset that performs well in differentiating PPD from HP, with greater accuracy. It can be noticed from the Tab. 6. that the HYBRID data set achieved better accuracy of more than 97 % with all the quartile features compared to the other datasets of the study and also achieved remarkable accuracy

of 98.6 % with Q2 features generated using RF feature selection and RF classification algorithms.

Tab. 5. Performance of the ML algorithms on various datasets with different feature sets

Feature set	Feature Selection Algorithm	CDT				HVLT				PDST				HYBRID			
		Accuracy of CA (%)				Accuracy of CA (%)				Accuracy of CA (%)				Accuracy of CA (%)			
		RF	NB	SVM	LR	RF	NB	SVM	LR	RF	NB	SVM	LR	RF	NB	SVM	LR
Q1	US	0.773	0.618	0.773	0.773	0.626	0.659	0.622	0.649	0.704	0.676	0.704	0.704	0.905	0.822	0.841	0.840
	RFE	0.783	0.536	0.793	0.793	0.741	0.626	0.723	0.626	0.746	0.662	0.732	0.732	0.958	0.829	0.849	0.832
	RFA	0.587	0.577	0.587	0.587	0.622	0.622	0.654	0.631	0.577	0.633	0.647	0.647	0.920	0.717	0.728	0.722
	RF	0.773	0.618	0.773	0.773	0.741	0.626	0.723	0.626	0.704	0.676	0.732	0.732	0.955	0.829	0.849	0.835
	UNION	0.793	0.567	0.783	0.793	0.723	0.640	0.705	0.672	0.732	0.704	0.718	0.718	0.982	0.805	0.841	0.832
	Mean	0.773	0.577	0.814	0.814	0.589	0.645	0.654	0.640	0.549	0.577	0.507	0.507	0.904	0.822	0.841	0.840
	Wt. Mean	0.773	0.567	0.814	0.814	0.566	0.645	0.649	0.640	0.704	0.676	0.704	0.704	0.905	0.819	0.849	0.835
Q2	US	0.762	0.567	0.721	0.721	0.658	0.654	0.663	0.668	0.690	0.704	0.676	0.676	0.947	0.799	0.837	0.832
	RFE	0.773	0.608	0.773	0.773	0.783	0.626	0.760	0.668	0.690	0.676	0.732	0.732	0.974	0.828	0.846	0.837
	RFA	0.597	0.587	0.608	0.608	0.695	0.585	0.705	0.631	0.661	0.647	0.690	0.690	0.976	0.795	0.837	0.831
	RF	0.783	0.608	0.773	0.773	0.764	0.612	0.774	0.622	0.718	0.633	0.732	0.732	0.986	0.814	0.843	0.840
	UNION	0.773	0.587	0.814	0.814	0.755	0.617	0.751	0.668	0.746	0.704	0.760	0.760	0.980	0.798	0.838	0.835
	Mean	0.773	0.577	0.814	0.814	0.686	0.668	0.677	0.645	0.619	0.662	0.647	0.647	0.949	0.802	0.838	0.832
	Wt. Mean	0.773	0.567	0.814	0.814	0.732	0.663	0.723	0.672	0.690	0.704	0.676	0.732	0.958	0.796	0.837	0.831
Q3	US	0.752	0.587	0.783	0.783	0.691	0.622	0.668	0.663	0.690	0.704	0.676	0.676	0.968	0.798	0.841	0.837
	RFE	0.783	0.587	0.814	0.814	0.751	0.635	0.741	0.677	0.732	0.676	0.690	0.690	0.979	0.814	0.838	0.840
	RFA	0.577	0.567	0.618	0.618	0.723	0.603	0.700	0.645	0.690	0.647	0.690	0.690	0.980	0.789	0.835	0.837
	RF	0.773	0.577	0.814	0.814	0.723	0.594	0.723	0.663	0.732	0.676	0.690	0.690	0.974	0.798	0.837	0.844
	UNION	0.762	0.577	0.783	0.783	0.755	0.566	0.728	0.672	0.732	0.704	0.732	0.732	0.973	0.798	0.835	0.835
	Mean	0.773	0.577	0.814	0.814	0.714	0.631	0.686	0.668	0.718	0.647	0.690	0.690	0.974	0.798	0.843	0.837
	Wt. Mean	0.773	0.567	0.814	0.814	0.723	0.603	0.709	0.668	0.704	0.704	0.732	0.676	0.979	0.798	0.838	0.841
Q4		0.762	0.577	0.783	0.783	0.741	0.576	0.714	0.668	0.718	0.704	0.760	0.760	0.973	0.798	0.835	0.835

*Yellow colored areas represent the highest accuracy obtained by CA in the respective Quartile feature set;

Tab. 6. The best performances of CA with the best FS algorithm with quartile set of features

Features	Analysis	CDT				HVLT				PDST				HYBRID			
		CA				CA				CA				CA			
		RF	NB	SVM	LR	RF	NB	SVM	LR	RF	NB	SVM	LR	RF	NB	SVM	LR
Q1	Acc	0.793	0.618	0.814	0.814	0.741	0.659	0.723	0.672	0.746	0.704	0.732	0.732	0.982	0.829	0.849	0.840
	FSA	UNION	US	Mean	Mean	RFE	US	RFE	UNION	RFE	UNION	RF	RF	UNION	RFE	RF	US
	Acc		0.618	0.814	0.814	0.741									0.829	0.849	0.840
	FSA		RF	Wt-Mean	Wt-Mean										RF	Wt-Mean	Mean
Q2	Acc	0.783	0.608	0.814	0.814	0.783	0.668	0.774	0.672	0.746	0.704	0.760	0.760	0.986	0.828	0.846	0.840
	FSA	RF	RFE	UNION	UNION	RFE	Mean	RF	Wt-Mean	UNION	US	UNION	UNION	RF	RFE	RFE	RF
	Acc		0.608	0.814	0.814						0.704						
	FSA		RF	Mean	Mean						UNION						
	Acc			0.814	0.814						0.704						
	FSA			Wt-Mean	Wt-Mean						Wt-Mean						
Q3	Acc	0.783	0.587	0.814	0.814	0.755	0.635	0.741	0.677	0.732	0.704	0.732	0.732	0.980	0.814	0.843	0.844
	FSA	RFE	US	RFE	RFE	UNION	RFE	RFE	RFE	RFE	US	Wt-Mean	UNION	RFA	RFE	Mean	RF
	Acc		0.587	0.814	0.814					0.732	0.704						
	FSA		RFE	RF	RF					RF	UNION						
	Acc			0.814	0.814					0.732	0.704						
	FSA			Mean	Mean					UNION	Wt-Mean						
	Acc			0.814	0.814												
	FSA			Wt-Mean	Wt-Mean												
Q4	Acc				0.783	0.741							0.760	0.973			

*Wt-Mean = Weighted Mean; FSA = Feature Selection Algorithm; Acc=Accuracy (in percentage); yellow colored areas represent the best accuracy obtained by CA in the respective Quartile feature set using the best FS mechanism;

Table 6. shows the performance metrics employed in this study to identify the behaviour of different CA working on different datasets while Tab. 7. shows the training response times involved. The lowest training response time of 8sec was possible with RF on HVLT dataset while LR achieved the highest training response time of 3600sec on the same dataset. Even though the HYBRID dataset achieved the highest training response time of 1620sec (approximately 27 minutes) with RF classification algorithm compared to other algorithms, the remarkable performance of 98.6 %, achieved using only non-motor symptoms places it

in the first position rather than the expensive clinical tests and time involved in the diagnosis process.

Tab. 7. Performance comparison with other datasets

Data Type / features quantity	Q1			Q2			Q3			Q4	
	Acc	FSA	CA	Acc	FSA	CA	Acc	FSA	CA	Acc	CA
CDT	0.814	Mean, Wt-Mean	LR, SVM	0.814	Mean	LR	0.814	Mean	LR	0.783	LR
HVLT	0.741	RF, RFE	RF	0.783	RFE	RF	0.755	UNION	RF	0.741	RF
PDST	0.746	RFE	RF	0.760	UNION	SVM, LR	0.732	RFE, RF, UNION, Mean, Wt-Mean	RF, SVM, LR	0.760	LR
HYBRID	0.982	UNION	RF	0.986	RF	RF	0.980	RFA	RF	0.973	RF

*Acc-Accuracy; FSA-Feature Selection Algorithm; CA-Classification Algorithm;

Tab. 8. Performance metrics of HYBRID dataset with different quartiles of features

Dataset	FS	FSA	CA	Acc	BA	P	R	F
HYBRID	Q1	UNION	RF	0.982	0.972	0.995	0.946	0.970
	Q2	RF		0.986	0.978	1	0.956	0.977
	Q3	RFA		0.980	0.959	1	0.917	0.957
	Q4	--		0.973	0.956	1	0.912	0.954

*FS-Feature Set; Acc-Accuracy; BA-Balanced Accuracy; P-Precision; R-Recall; F-F1 Score

Tab. 9. Training response Time of ML algorithms on datasets

Dataset	RF	NB	SVM	LR
CDT	480	120	23	20
HVLT	8	240	120	3600
PDST	2520	240	33	32
HYBRID	1620	240	41	25

*Response Time in seconds

Tab. 10. Performance comparison with existing works

Ref	Count	Dataset used	Acc(%)
[24]	PPD:423; HC:195	RBDSQ and UPSIT	85.48
[25]	PPD:342; HC:157 SWEDD:51	Features derived from SPECT SCAN images	82.2
[34]	PPD:109; HC:40	audio	91.83
[35]	PPD:10; HC:8	Genetic, audio	88.9 91.83
[41]	Voice:5826; Demographics: 6805	Voice and demographics	85
[42]	PPD:401; HC:183	Rapid Eye Movement, olfactory loss, Cerebrospinal fluid data, and dopaminergic imaging markers	96.45
[43]	PPD:29; HC:18	Gait	95
[48]	Biocruces:96 PPMI:687	SDMT, BJLOT, MoCA, HVLT, BSIT, GDS and autonomic manifestations	86.3
Proposed	PPD:106; HC:106	SLEEP, HVLT, CDT, HYBRID	98.65

*RBDSQ-Rapid eye movement sleep Behaviour Disorder Screening Questionnaire; UPSIT-University of Pennsylvania Smell Identification Test; SDMT-Symbol Digit Modalities Test; BJLOT- Benton Judgment of Line Orientation Test; MoCA- Montreal Cognitive Assessment; BSIT- Brief Smell Identification Test; GDS-Geriatric Depression Scale; MLP-Multi Layer Perceptron; Acc-Accuracy;

4.2. Performance Evaluation

Tab. 11. Interpretability of HYBRID dataset vs other datasets

	Features	**HYBRID accuracy (hb)**	**CDT accuracy ©**	**HVLT accuracy (h)**	**PDST accuracy (p)**
	Q1	0.982	0.793	0.741	0.746
	Q2	0.986	0.783	0.783	0.746
	Q3	0.980	0.783	0.755	0.732
	Q4	0.973	0.762	0.741	0.718
	BCT for HYBRID and CDT				
	(hb-c)	$\overline{d}$	s_d^2	PSD	μ_d
Q1	0.189	0.200	0.000087	0.00465	43.01
Q2	0.203				
Q3	0.197				
Q4	0.211				
	BCT for HYBRID and HVLT				
	(hb-h)	$\overline{d}$	s_d^2	PSD	μ_d
Q1	0.241	0.225	0.000263	0.00811	27.75
Q2	0.203				
Q3	0.225				
Q4	0.232				
	BCT for HYBRID and PDST				
	(hb-p)	$\overline{d}$	s_d^2	PSD	μ_d
Q1	0.236	0.245	0.000072	0.00423	57.94
Q2	0.240				
Q3	0.248				
Q4	0.255				

BCT (Corani & Benavoli, 2015) is a statistical technique to compare the performance of two ML algorithms. It uses bayesian inference to get a probabilistic evaluation of which algorithm performs better than the other, using the correlation. For each dataset, the process involves identifying the performance metrics differences between the two algorithms, modelling these differences with a Bayesian framework with presumptive prior distributions, then updating these priors with observed data to produce a posterior distribution. Next, posterior distribution is assessed to find the likelihood that one approach performs better than the other. Finally, Cumulative Distributive Function (CDF), which gives the cumulative probability up to a certain point. In the proposed study, as HYBRID dataset achieved the best performance with all quartiles of features using RF CA, we compared the performance of RF algorithm with all the datasets using BCT as shown in Tab. 10. The Probability that $P(\mu_d > 0|data \approx CDF(\mu|\square|d)$is 1, as the $\mu_d$values are very large and CDF will be obviously 1, indicating that the HYBRID dataset performs well, compared to other data sets.

5. DISCUSSION

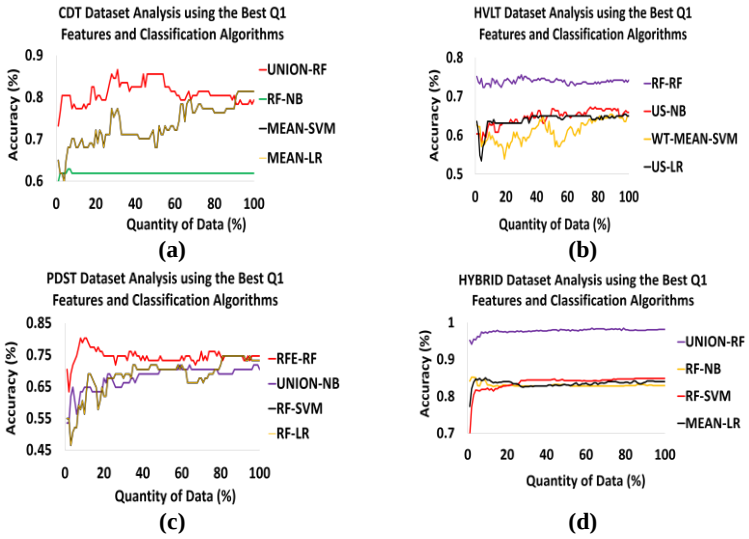

Fig. 3. (a), (b), (c), (d): CDT, HVLT, PDST and HYBRID datasets performance with the best Q1 features respectively

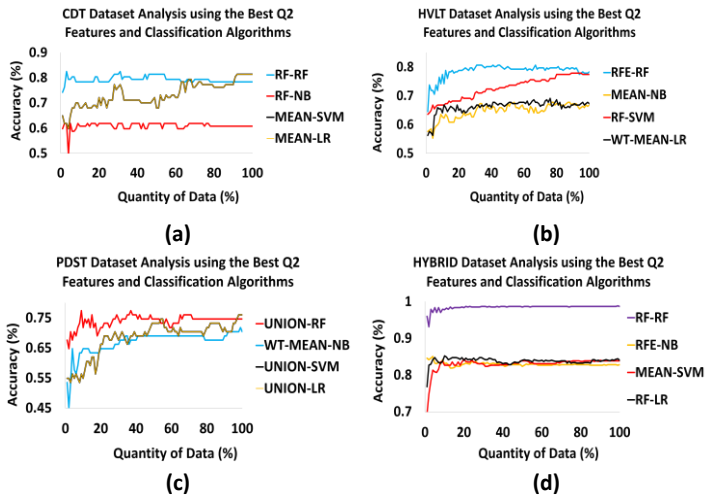

Fig. 4. (a), (b), (c), (d): CDT, HVLT, PDST and HYBRID datasets performance with the best Q2 features respectively

It can be observed that the RF classification algorithm performed well for HVLT, PDST, HYBRID datasets using RFE feature set (Fig. 3(b), 3(c) and Fig. 3(d)) while SVM and LR outperformed for CDT data set using mean feature set with Q1 features (Fig. 3(a)). RF classification algorithm performed well for HVLT and HYBRID datasets using RFE and RF (Fig. 4(b) and Fig. 4(d)) feature sets respectively, for PDST data set using UNION feature

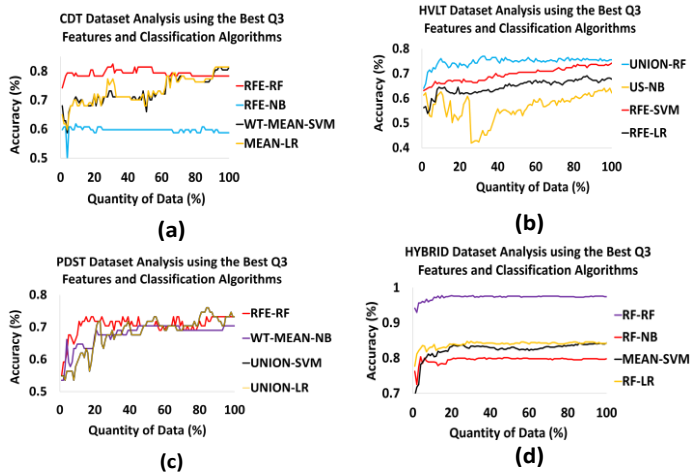

Fig. 5. (a), (b), (c), (d): CDT, HVLT, PDST and HYBRID datasets performance with the best Q3 features respectively

RF classification algorithm performed well for HVLT and HYBRID datasets using RFE and RF (Fig. 4(b) and Fig. 4(d)) feature sets respectively, for PDST data set using UNION feature set with Q2 features (Fig. 4(c)). SVM and LR outperformed for CDT data set using mean feature set with Q2 features (Fig. 4(a)). With Q3 feature set, it can be noticed that SVM and LR outperformed for CDT dataset using mean and weighted mean feature sets (Fig. 5(a)) while RF, SVM and LR outperformed for PDST data set using RFE and UNION feature sets (Fig. 5(c)). RF outperformed HVLT and HYBRID datasets using features obtained using RF (Fig. 5(b) and 5(d)).

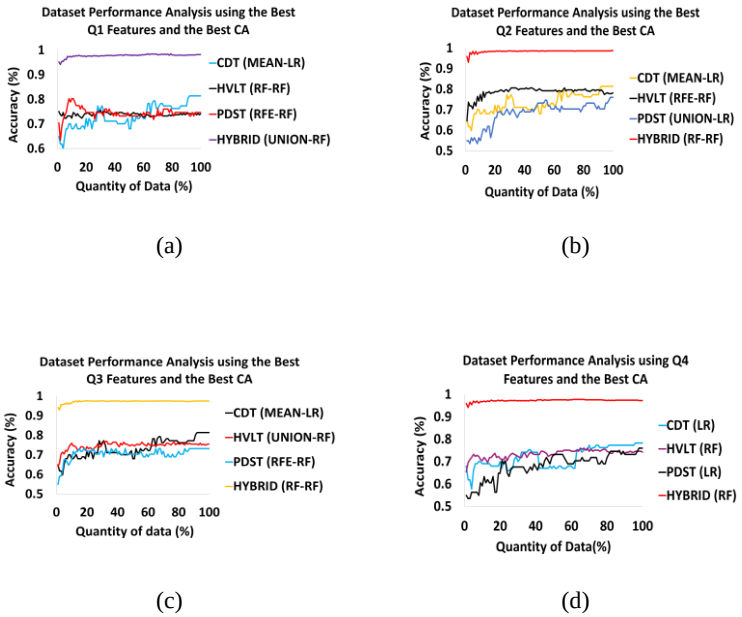

Fig. 6. (a), (b), (c), (d): datasets performance with the best CA and Q1, Q2, Q3 and Q4 features respectively

LR outperformed for all the data sets, without using any feature selection mechanism (Fig. 6(a), 6(b), 6(c) and 6(d)). All the algorithms obtained an accuracy of above 70 % in discriminating PPD from HP (Tab. 5.) using Q3 feature sets generated by the discussed feature selection mechanisms, which proves that many learning strategies achieve quite good detection rates. Remarkably, RF classification algorithm obtained an accuracy of over 97 % on all quantitative feature sets using HYBRID dataset. The results clearly show that the custom-built HYBRID dataset outperformed the remaining data sets. The performance improvement provided by the feature selection process was especially remarkable for Q2 feature set of the proposed HYBRID dataset yielding an accuracy of 98.6 % with RF algorithm being used for both feature selection and classification purpose.

From Tab. 6, it can be observed that the accuracies increased from Q1 to Q2 while dropped from Q2 to Q3 and further dropped to Q4 using HVLT and HYBRID datasets, while it is maintained constant in all quartiles for CDT dataset. It can also be noted that the accuracy is increased slightly from Q3 to Q4 equating with Q2 in case of PDST data set. This clearly shows that the quantitative feature selection mechanism plays a vital role on the CA performance.

From Tab. 7, it can be observed that a precision value of '1' was achieved in Q2, Q3 and Q4 quartiles, which is desirable in medical diagnosis. A perfect precision indicates that the model has high confidence in its positive predictions, proving that the disease can be diagnosed truly. The model is very conservative, only predicting positive when it is very sure, hence avoiding any incorrect positive predictions. The highest recall value of 0.956 in Q2 set indicates strong performance in identifying TP. The highest F1- score value of 0.977 can also be observed in Q2 set which indicates that it is the best performing quartile overall. The consistent closeness of accuracy and balanced accuracy values indicates that the model's performance is balanced across classes, avoiding significant biases. Q3 and Q4 have the lowest Balanced Accuracy values of 0.959 and 0.956, respectively, suggesting that the performance on one or both classes is slightly less balanced compared to Q1 and Q2. It is evident that Q2 feature set of HYBRID dataset is the best performing quartile with the highest Accuracy and Balanced Accuracy, indicating a well-balanced and highly effective model that can be used to differentiate PPD from HP. Balanced accuracy is also highest for Q2 and lowest for Q4, indicating that the classifier is well-balanced and performs consistently across classes, with Q2 being the most balanced. Precision is perfect for Q2, Q3, and Q4, indicating no false positives in these cases, while for Q1, precision is slightly lower, suggesting a minimal number of false positives. Recall is highest for Q2 and lowest for Q4, indicating that the classifier's ability to identify true positives was best in Q2 and weakest in Q4. F-Score, which is the harmonic mean of precision and recall, is highest for Q2 and lowest for Q4, highlighting that Q2 had the best balance of precision and recall. Overall, the classifier performed best with the features in Q2, as evidenced by the highest accuracy, balanced accuracy, recall, and F1-score, allowing the classifier to identify true positives and true negatives most effectively. The precision remains consistently high across Q2, Q3, and Q4, indicating that the classifier is effective at avoiding false positives with these feature sets. Recall shows more variability, suggesting that while the classifier can avoid false positives well, its ability to identify all true positives depends more on the feature set used. The feature set in Q4, while still performing well, shows the lowest values across all metrics compared to Q1, Q2, and Q3, indicating a possible trade-off between various datasets using ML algorithms. Considering the performance metrics, Q2 seems to be the

most effective feature set for this classifier, and it might be beneficial to analyse and understand why Q2 features perform best to potentially apply similar selection criteria to other datasets. It would also be useful to examine the specific features in Q2 that contribute to its high performance and compare them with those in Q3 and Q4 to optimize feature selection further. Despite the strong performance, fine-tuning the model further with hyperparameter optimization, and potentially incorporating other ML models could improve the results even more.

Table 8. gives the training response times involved in training the discussed datasets using the RF, NB, SVM and LR classification models. It is evident that HVLT dataset using RF classification model was trained consuming the lowest time (8 seconds) while the highest time was required to train using LR model (3600 seconds).

Table 9. Shows the performance of the proposed model compared with the existing works and it can be noticed that the proposed model stands in the first position with an accuracy of 98.65 %.

Large evidence is provided with BCT that the HYBRID dataset significantly outperforms the CDT, HVLT, and PDST datasets across the 4 quartiles feature sets, using RF classification method, which can be observed from Tab.10. The mean difference in accuracy is substantial with a P (μ_d>0|data) of virtually 1, indicating very high confidence in these results.

6. CONCLUSION

The study employed various validation techniques, utilizing 10 runs of a 10-fold cross-validation method. Evaluation primarily focused on accuracy metrics, complemented by analysis of F1-score, Precision, and Recall. The Bayesian approach interprets probability subjectively, viewing it as a measure of belief in the face of uncertainty. The approach was used to determine the best-performing model among comparisons involving 7 distinct feature selection mechanisms and 4 trained models on 4 different datasets. Evaluation metrics included confusion matrix, accuracy, balanced accuracy, specificity, precision, recall and f1-score. Currently, there are limited studies employing ML algorithms to detect PD based on non-motor symptoms, often supported by more complex data such as clinical images and biofluid biomarkers. Previous research focused on using individual non-motor features to distinguish PPD from HP with various objectives. In the proposed study, a strong PD detection rate of 97.3% was achieved without employing any feature selection mechanisms, and up to 98.6% using second quartile features. This was accomplished through a hierarchical screening strategy where level-1 identified optimal feature selection mechanisms, level-2 selected the best CA paired with these mechanisms, and level-3 identified the best dataset resulting in the highest predictive accuracy. HVLT and PDST datasets performed the best way with RFE and UNION FS methods across all quartile features, CDT dataset excelled with Mean feature selection, and the HYBRID dataset optimized with RF, RFA, and UNION feature selection methods for Q1, Q2, and Q3 features respectively at level-1. RF and LR algorithms were identified as optimal at level-2. Ultimately, the HYBRID dataset comprising HVLT, PDST, and CDT datasets achieved a notable 98.6% accuracy using second quartile features selected by RF, and classified by RF algorithm. The authors assert that this methodological hierarchy can significantly enhance

early detection of PD. The BCT conducted on RF performance evaluation on various datasets provides a valuable insight for researchers and clinicians for clinical use of the proposed work. A primary limitation of the proposed study was its focus solely on classifying HP and PPD using existing numerical data related to non-motor symptoms. Assessment of the disease severity was not included. Other important non-motor assessments tests such as UPSIT, BJLOT, GDS, Wechsler Adult Intelligence Scale (WAIS), PD Anxiety Scale Test (PDAST), Boston Naming Test (BNT) were not included in the study. Another significant limitation was to incorporate imaging data and compare with the numerical data in analysing the performance. Thus, the authors encourage efforts aimed at acquiring additional data from both PPD and HP. Future studies should aim to include different imaging data to enhance differentiating PPD from HP accurately.

Ethical Approval

This manuscript reports studies which does not involve human participants/data/tissue and animals.

Conflict of Interest

The authors have no conflict of interests to declare that are relevant to the content of this article.

Authors Contributions

Anitha Rani Palakayala contributed to conceptualization, methodology, software, data curation, formal analysis, writing—original draft. Kuppusamy P contributed to conceptualization, methodology, supervision, result analysis, writing—review and editing final draft and validation.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Availability of data and materials

Data used in this study is available from PPMI public database and can be downloaded from the web site: https://www.ppmi-info.org/data.

REFERENCES

Adeli, E., Shi, F., An, L., Wee, C.-Y., Wu, G., Wang, T., & Shen, D. (2016). Joint feature-sample selection and robust diagnosis of Parkinson's disease from MRI data. *NeuroImage, 141*, 206-219. https://doi.org/10.1016/j.neuroimage.2016.05.054

Ali, L., Chakraborty, C., He, Z., Cao, W., Imrana, Y., & Rodrigues, J. J. P. C. (2022). A novel sample and feature dependent ensemble approach for Parkinson's disease detection. *Neural Computing and Applications*, *35*, 15997–16010. https://doi.org/10.1007/s00521-022-07046-2

Alkhatib, R., Diab, M. O., Corbier, C., & Badaoui. M. E. (2020). Machine Learning algorithm for gait analysis and classification on early detection of Parkinson. *IEEE Sensors Letters*, *4*(6), 1-4. https://doi.org/10.1109/LSENS.2020.2994938

Armañanzas, R., Bielza, C., Chaudhuri, K. R., Martinez-Martin, P., & Larrañaga, P. (2013). Unveiling relevant non-motor Parkinson's disease severity symptoms using a machine learning approach. *Artificial Intelligence in Medicine*, *58*(3), 195-202. https://doi.org/10.1016/j.artmed.2013.04.002

Benedict, R. H. B., Schretlen, D., Groninger, L., & Brandt, J. (1998). Hopkins verbal learning test - Revised: Normative data and analysis of inter-form and test-retest reliability. *Clinical Neuropsychologist*, *12*(1), 43-55. https://doi.org/10.1076/clin.12.1.43.1726

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32.

Chaudhuri, K. R., Pal, S., DiMarco, A., Whately-Smith, S., Bridgman, K., Mathew, R., Pezzela, F. R., Forbes, A., Högl, B., & Trenkwalder, C. (2002). The Parkinson's disease sleep scale: a new instrument for assessing sleep and nocturnal disability in Parkinson's disease. *J Neurol Neurosurg Psychiatry*, *73*(6), 629-635. https://doi.org/10.1136/jnnp.73.6.629

Connolly, B. S., & Lang, A. E. (2014). Pharmacological treatment of Parkinson disease: a review. *JAMA*, *311*(16), 1670–1683. https://doi.org/10.1001/jama.2014.3654

Corani, G., & Benavoli, A. (2015). A bayesian approach for comparing cross-validated algorithms on multiple data sets. *Machine Learning*, *100*, 285-304. https://doi.org/10.1007/s10994-015-5486-z

Cordella, F., Paffi, A., & Pallotti, A. (2021). Classification-based screening of Parkinson's disease patients through voice signal. *2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA)* (pp. 1-6). IEEE. https://doi.org/10.1109/MeMeA52024.2021.9478683

De Lau, L. M. L., & Breteler, M. M. B (2006). Epidemiology of Parkinson's disease. *The Lancet Neurology*, *5*(6), 525-535. https://doi.org/10.1016/S1474-4422(06)70471-9

Drotár, P., Mekyska, J., Rektorová, I., Masarová, L., Smékal, Z., & Faundez-Zanuy, M. (2014). Decision support framework for Parkinson's disease based on novel handwriting markers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *23*(3), 508-516. https://doi.org/10.1109/tnsre.2014.2359997

Govindu, A., & Palwe, S. (2023). Early detection of Parkinson's disease using machine learning. *Procedia Computer Science*, *218*, 249-261. https://doi.org/10.1016/j.procs.2023.01.007

Gunakala, A., & Shahid, A. H. (2023). A comparative study on performance of basic and ensemble classifiers with various datasets. *Applied Computer Science*, *19*(1), 107-132. https://doi.org/10.35784/acs-2023-08

Haq, A. U., Li, J. P., Memon, M. H., Khan, J., Malik, A., Ahmad, T., Ali, A., Nazir, S., Ahad, I., & Shahid, M. (2019). Feature selection based on L1-Norm support vector machine and effective recognition system for Parkinson's Disease using voice recordings. *IEEE Access*, 7, 37718-37734. https://doi.org/10.1109/ACCESS.2019.2906350

Hosmer, D. W., Lemeshow, S. H., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons.

Huang, F., Xu, H., Shen, T., & Jin, L. (2021). Recognition of Parkinson's Disease based on residual Neural Network and voice diagnosis. *2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)* (pp. 381-386). IEEE. http://dx.doi.org/10.1109/ITNEC52019.2021.9586915

Mabrouk, R., Chikhaoui, B., & Bentabet, L. (2018). Machine learning based classification using clinical and DaTSCAN SPECT imaging features: a study on Parkinson's disease and SWEDD. *IEEE Transactions on Radiation and Plasma Medical Sciences*, *3*(2), 170-177. https://doi.org/10.1109/TRPMS.2018.2877754

Mainland, B. J., & Shulman, K. I. (2017). Clock drawing test. In A. J. Larner (Ed.), *Cognitive Screening Instruments* (pp. 67–108). Springer International Publishing. https://doi.org/10.1007/978-3-319-44775-9_5

Martinez-Eguiluz, M., Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Perona, I., Murueta-Goyena, A., Acera, M., Del Pino, R., Tijero, B., Gomez-Esteban, J. C., & Gabilondo, I. (2023). Diagnostic classification of Parkinson's disease based on non-motor manifestations and machine learning strategies. *Neural Computing and Applications*, *35*, 5603-5617. https://doi.org/10.1007/s00521-022-07256-8

Mei, J., Desrosiers, C., & Frasnelli, J. (2021). Machine Learning for the diagnosis of Parkinson's disease: A review of literature. *Frontiers in Aging Neuroscience*, *13*, 633752. https://doi.org/10.3389/fnagi.2021.633752

Moradi, S., Tapak, L., & Afshar, S. (2022). Identification of novel non invasive diagnostics biomarkers in the Parkinson's diseases and improving the disease classification using support vector machine. *BioMed Research International*, *2022*(1), 009892. https://doi.org/10.1155/2022/5009892

Nuvoli, S., Spanu, A., Fravolini, M. L., Bianconi, F., Cascianelli, S., Madeddu, G., & Palumbo, B. (2020). [123i] Metaiodobenzylguanidine (MIBG) cardiac scintigraphy and automated classification techniques in Parkinsonian disorders. *Molecular Imaging and Biology*, *22*(3), 703-710. https://doi.org/10.1007/s11307-019-01406-6

Pahwa, R., & Lyon, K. E. (2010). Early diagnosis of Parkinson's disease: recommendations from diagnostic clinical guidelines. *The American Journal Managed Care*, *16*, 94-99.

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. *In: Machine Learning* (pp. 101-121). Elsevier. http://dx.doi.org/10.1016/B978-0-12-815739-8.00006-7

Prashanth, R., Roy, S. D., Mandal, P. K., & Ghosh, S. (2014). Parkinson's disease detection using olfactory loss and REM sleep disorder features. *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 5764-5767). IEEE. https://doi.org/10.1109/embc.2014.6944937

Raundale, P., Thosar, C., & Rane, S. (2021). Prediction of Parkinson's disease and severity of the disease using Machine Learning and Deep Learning algorithm. *2021 2nd International Conference for Emerging Technology (INCET)* (pp. 1-5). IEEE. https://doi.org/10.1109/INCET51464.2021.9456292

Ricciardi, C., Amboni, M., De Santis, C., Ricciardelli, G., Improta, G., D'Addio, G., Cuoco, S., Picillo, M., Barone, P., & Cesarelli, M. (2020). Machine learning can detect the presence of mild cognitive impairment in patients affected by Parkinson's disease. *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)* (pp. 1-6). IEEE. https://doi.org/10.1109/MeMeA49120.2020.9137301

Sakar, B. E., Isenkul M. E., Sakar, C. O., Sertbas, A., Gurgen, F., Delil, S., Apaydin, H., & Kursun, O. (2013). Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE J Biomed Health Informatics*, *17*(4), 828-834. https://doi.org/10.1109/jbhi.2013.2245674

Schrag, A., Jahanshahi, M., & Quinn, N. (2000). How does Parkinson's disease affect quality of life? A comparison with quality of life in the general population. *Movement Disorders*, *15*(6), 1112-1118. https://doi.org/10.1002/1531-8257(200011)15:6%3C1112::aid-mds1008%3E3.0.co;2-a

Smyth, C., Anjum, M. F., Ravi, S., Denison, T., Starr, P., & Little, S. (2023). Adaptive deep brain stimulation for sleep stage targeting in Parkinson's disease. *Brain Stimulation*, *16*(5), 1292-1296. https://doi.org/10.1016/j.brs.2023.08.006

Thangaleela, S., Sivamaruthi, B. S., Kesika, P., Mariappan, S., Rashmi, S., Choeisoongnern, T., Sittiprapaporn, P., & Chaiyasut, C. (2023). Neurological insights into sleep disorders in Parkinson's disease. *Brain Sciences*, *13*(8), 1202. https://doi.org/10.3390/brainsci13081202

Trenkwalder, C., Kohnen, R., Högl, B., Metta, V., Sixel-Döring, F., Frauscher, B., Hülsmann, J., Martinez-Martin, P., & Chaudhuri, K. R. (2011). Parkinson's disease sleep scale-validation of the revised version PDSS-2. *Movement Disorders*, *26*(4), 644-652. https://doi.org/10.1002/mds.23476

Vellido, A. (2020). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, *32*, 18069-18083. https://doi.org/10.1007/s00521-019-04051-w

Wang, W., Lee, J., Harrou, F., & Sun, Y. (2020). Early detection of Parkinson's disease using Deep Learning and Machine Learning. *IEEE Access*, *8*, 147635-147646. https://doi.org/10.1109/ACCESS.2020.3016062

Wodzinski, M., Skalski, A., Hemmerling, D., Orozco-Arroyave, J. R., & Nöth, E. (2019). Deep Learning approach to Parkinson's disease detection using voice recordings and convolutional Neural Network dedicated to image classification. *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 717-720). IEEE. https://doi.org/10.1109/EMBC.2019.8856972

Wroge, T. J., Özkanca, Y., Demiroglu, C., Si, D., Atkins, D. C., & Ghomi, R. H. (2018). Parkinson's disease diagnosis using Machine Learning and voice. *2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)* (pp. 1-7). IEEE. https://doi.org/10.1109/SPMB.2018.8615607

Yu, T., & Zhu, H. (2020). Hyper-parameter optimization: a review of algorithms and applications. *ArXiv, abs/2003.05689*. https://doi.org/10.48550/arXiv.2003.05689

Zhang, H. (2004). The optimality of naive bayes.*The Florida AI Research Society*.