JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 , Special Issue, No.-11, May (2024) pp 133-146 ISSN (Print) 0973-8975

FEATURE SELECTION USING EXTRA TREES CLASSIFIER FOR PARKINSON'S DISEASE CLASSIFICATION

Gauri Sabherwal[1], Amandeep Kaur[2], Uday Malhotra[3]

[1]Department of Computer Science and Engineering, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India

[2]Department of Computer Science and Engineering, Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India

[3]Department of Computer Science and Engineering, Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India

Email : gauri.cse.21@gmail.com[1], amandeep@chitkara.edu.in[2], udaymalhotra04@gmail.com[3]

Corresponding Author: **Gauri Sabherwal**

https://doi.org/10.26782/jmcms.spl.11/2024.05.00010

(Received: March 14, 2024; Revised: April 28, 2024; Accepted: May 16, 2024)

Abstract

Parkinson's disease (PD) is chronic, permanent, and life-threatening. Neurologically protective treatments for PD rely on early detection. Recent studies have demonstrated that clinical data, cerebrospinal Fluid (CSF) based proteomes, and gene mutations are important biomarkers for accurate and early detection of PD. This study aims to investigate the heterogeneous data comprised of CSF-based clinical data, CSF-based proteomic analysis data as well as the mutation information of the genes, Glucose Beta Acid (GBA), leucine-rich kinase (LRRK2) to classify controls into PD-affected and Healthy Control (HC). The dataset contains 1103 controls (569 PD affected and 534 HC). Automated Machine Learning (AutoML) framework using PyCaret is utilized. The study has proposed an Extra Tree Classifier (ETC) as a feature selection mechanism to select features that significantly affect the PD classification. Selected features are further used to train Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT) classifiers. Accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUC-ROC), and the confusion matrix are used to evaluate the performance of classifiers. RF has depicted the best performance in terms of accuracy value of 96.12%, sensitivity of 95.59%, and specificity of 95.34% while LR has shown the highest AUC value of 98.33. RF has made the highest number of correct predictions 316 out of 331.

Gauri Sabherwal et al

***A Special Issue on* 'Recent Evolution in Applied Sciences and Engineering'.**

Keywords: Parkinson's Disease, CSF, Feature Selection, Extra Tree Classifier, Machine Learning; Random Forest; Logistic Regression.

I. Introduction

PD is a neurodegenerative condition that ranks second in prevalence to Alzheimer and its global prevalence is increasing as the world's population ages. The cause of this neurological disorder is a dopamine deficiency in the brain area. Since dopamine is a neurochemical messenger, it makes it easier for impulses to reach the basal ganglia, which are essential for controlling movement and synchronization. The reduction in dopamine levels arises from the dysfunction or death of basal ganglia cells [II]. This illness presents with symptoms such as paralysis of the muscles, rigidity of the limbs, difficulty speaking or writing, and problems with standing up, balancing, and movement. Researchers have been studying PD exhaustively for decades, but still don't know what causes the death of dopamine-producing cells [XXI].

The best way to improve a patient's quality of life is to catch the condition early. Early diagnosis makes managing the PD symptoms easier and prevents the condition from worsening faster. [V]. The majority of PD cases are recognized as unexplained; however, an increasing number of genetic changes either expedite the progression of PD or raise the risk. In the contemporary period, Machine Learning (ML) and Deep Learning (DL) have sparked a revolutionary change in the healthcare sector, resulting in notable breakthroughs in diagnoses and treatment approaches [VIII, XIX, VI, XVIII].

A growing amount of research suggests that the pathophysiology of PD involves both genetic and evolving immune reactions [XXVII]. One of the most prevalent PD risk genes is GBA and another is LRRK2[III]. In a recent study, PD pathogenesis and clinical heterogeneity have been assessed from a molecular, pathological, and clinical standpoint. Analysis of proteomics based on CSF has created new avenues for examining this diversity[XII]. Another study has demonstrated that the CSF proteome has clinically significant findings about the onset and course of PD by combining ML and high-throughput proteomics [XXV].

CSF-based clinical data, CSF-based proteomic analysis data as well as genetic mutation information have opened the door to the discovery of putative biomarkers and improved understanding of the PD's history. The study proposes to utilize ETC for feature selection based on selecting the significant features. Further, three classifiers i.e. RF, DT, and LR are compared to classify controls as PD-affected controls and HC. The paper's structure is arranged in the following order: Previous studies on feature selection and the classification model utilized for PD diagnosis are presented in Section II. The research work's methodology is explained in Section III. Both the feature selection and classification results are examined in Section IV. Section V discusses the conclusions and future directions.

II. Related Work

Researchers have been conducting a lot of studies to examine the diagnosis of PD disease using different modalities such as speech, Motor data, gene data, CSF data, clinical data, handwriting patterns, etc. The XGB model has been put forth by Pantaleo et al. to identify PD patients. A total of 550 samples of gene expression data from PD and HC patients are used. This method yields an AUC of 0.72, indicating a moderate level of accuracy. Numerous putative genes and pathways have also been identified by the study for further investigation and validation [XII]. Markello et al. has developed a unified analysis approach for CSF, imaging, and clinical data [X]. Their results have demonstrated that this integrated strategy performed better in efficiently collecting commonalities across patients across modalities than traditional methods, which include merging data across dimensions. Oh et al. have presented a 13-layer Convolutional Neural Network (CNN) classifier to distinguish participants into PD and HC using resting-state EEGs obtained from 20 HC and 20 PD patients. [XI]. The suggested model has a sensitivity of 84.7%, specificity of 92%, and accuracy of 88.3%.

Rasheed et al. have developed a DL model for the diagnosis of fresh-onset PD using voice samples from 23 PD patients and 8 HC [XVII]. Principal Component Analysis (PCA) is employed for preprocessing the voice data. Using the 15 most discriminative traits, an accuracy rate of 97.5% has been achieved. Using a specifically created tele-diagnosis technology, a research study [XX] has investigated the feasibility of remotely diagnosing PD. From the signals of both PD patients and healthy people, the algorithm has identified particular traits (features). These features have been used to train a Support Vector Machine (SVM) classifier. The goal of the study is to use the fewest features required to attain the maximum accuracy achievable.

A classifier has been proposed by Shetty et al. [XXIII] for PD classification. A dataset with gait characteristics from patients with Huntington's disease, amyotrophic lateral sclerosis, PD controls, and HC is employed. SVM classifier is used with a Gaussian RBF Kernel. Compared to other neurodegenerative diseases, SVM has been identified in 75% of PD patients. The reported false positive rate for Huntington's disease patients is 30.76%, while it is 0% for those with amyotrophic lateral sclerosis. Abdulhay et al. have suggested a way to use ML techniques based on the gait data to study gait and tremor. Using peak detection and pulse duration to extract several gait parameters, they have found 92.7% accuracy in diagnosing PD [I].

A collection of six distinct ML approaches has processed eighteen different configurations of gait characteristics derived from data about 25 controls. Classification accuracy for each configuration and ML technique is determined by majority vote. Additionally, two meta-classifiers with different weighting schemes have been incorporated based on individual technique's output. Classifiers' average classification accuracy has varied from 63% to 80%, and for one meta-classifier setup, it has reached 96% [IV]. In [XXVI], the authors have used a statistical pooling

technique to improve the dataset and ReliefF to identify the most important features. Their proposed Parkinson's diagnosis model has achieved the highest accuracies of 91.25% with SVM and 91.23% with KNN.

This study aims to utilize a heterogeneous dataset containing CSF-based clinical data, CSF-based proteomics analysis data, and gene mutation information for PD classification. The proposed approach is to use ETC for feature selection and to evaluate the performance of three ML models: RF, DT, and LR for classifying controls into PD-affected and HC.

III. Material and Method

Data Collection

In this study, we have used a dataset that is publicly available on the Parkinson's Progression Markers Initiative (PPMI) web page (https:// www.ppmi-info.org/access-data-specimens/download-data). The dataset is heterogeneous as it comprises CSF-based clinical data, CSF-based proteomic analysis data, and mutation status of the genes GBA and LRKK2. The dataset contains a total of 1103 samples from 569 PD affected and 534 HC. As far as we know, this is, to date, the largest CSF-based proteomic dataset with gene mutation information for classifying controls into PD-affected and HC. Table 1 summarizes the attributes of the dataset.

Table 1: Dataset Attributes

ATTRIBUTE	DESCRIPTION
Patno	Patient Identifier
SampleID	Sample Identifier
Mutation_Status	Mutation Status (GBA+ = carrier of GBA mutations N370S, L483P, L444P, IVS2+1 or 84GG; LRRK2+ = carrier of LRRK2 mutations G2019S or R1441G; noMut = non carrier of mutations in GBA, LRRK2 or SNCA)
Disease_Status	Disease Status (PD = Parkinson's Disease, HC = Control)
Gender	Gender
Age	Age at enrollment (years) = (ENROLLDT - BIRTHDT)
CNo	Study Center Identifier
LEDD_CAT	Levodopa Equivalent Daily Dose (true / false)
updrs1_score	Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) Part I Score
updrs2_score	MDS-UPDRS Part II Score
updrs3_score	MDS-UPDRS Part III Score
moca	Montreal Cognitive Assessment Total Score
mean_caudate	DaTSCAN mean bilateral caudate thickness (mm) = (CAUDATE_R + CAUDATE_L) / 2
mean_putamen	DaTSCAN mean bilateral putamen thickness (mm) = (PUTAMEN_R + PUTAMEN_L) / 2

mean_striatum	DaTSCAN mean bilateral striatum thickness (mm) = (CAUDATE_R + CAUDATE_L + PUTAMEN_R + PUTAMEN_L) / 4
Ptau	Phospho Tau 181P concentration in CSF (pg/mL)
Ttau	Total Tau concentration in CSF (pg/mL)
Asyn	Alpha Synuclein concentration in CSF (pg/mL)
Abeta	Amyloid beta 1-42 concentration in CSF (pg/mL)
PC1	Proteomic analysis using (SomaScan) Principal Component 1
PC2	Proteomic analysis using (SomaScan) Principal Component 2
PC3	Proteomic analysis using (SomaScan) Principal Component 3
PC4	Proteomic analysis using (SomaScan) Principal Component 4

A. Tools

The study has employed PyCaret, an AutoML open-source Python package that automatically compares the consistent top ML algorithms. PyCaret designs and establishes fast models and offers an optimal experience with a lesser number of lines of code and productivity with ML [XXIV]. It tests a wide range of models quickly, giving the data scientist a clear picture of which models perform well in both classification and regression tasks [XV].

B. Data Preprocessing

An important part of data processing is preprocessing, which helps the model to eliminate superfluous information. Using the Pandas package, the dataset has been imported as a CSV file into the Google Colab platform. We have checked the data for missing values. Mean imputation is used to fill in the dataset's missing values [XXIV]. The columns like Patno, SampleID, and CNo are dropped. The final data set used in the study includes 20 attributes and 1103 instances.

C. Feature Selection using extra trees classifier

ML relies heavily on feature selection to achieve optimal performance. We have employed ensemble ETC-based feature selection to select the significant features. To increase the accuracy of prediction and decrease overfitting, the ETC [XXII] fits randomized decision trees, or extra trees, on different subsets within the dataset. This process determines the significance of the features, which can then be utilized to filter out irrelevant data. The model's feature importance property is used to calculate feature importance. Data features are ranked between 0 and 1 based on their importance. A feature that scores higher is likely more relevant to the output variable. When building models, this score aids in selecting the most significant features and eliminating the least significant ones.

An ETC is initially initialized before selecting feature subsets. Meta estimators are built first by the classifier. Each estimator depicts the count of trees in the forest. When searching for the best tree split, the attribute max_features determines how many features to consider, and the search for a split doesn't stop until there is at least one valid partition of the node samples, even if it takes longer than max_features to

effectively examine the features. Each decision tree at each test node receives a random sample of n features from which it must choose the best feature for dividing the data using the Gini Index. By default, the function "Gini" uses the feature_importances_ argument to calculate the feature's Gini Importance. To choose significant aspects from the result, the average is utilized. While averaging is used to choose important characteristics from the output, the largest (n) function is used to choose the best n features.

D. Classification

The selected features are used for the classification of controls into PD-affected and HC. In this paper three classifiers: RF, LR, and DT are implemented, using Pycaret, an AutoML Python package. During training, the RF algorithm builds a diverse ensemble of decision trees. The final prediction for a new data point is made by selecting the class that receives the most votes from these individual trees [XIII]. LR is a statistical method for examining datasets in which the result is influenced by several independent variables. Through LR analysis, it is possible to determine the optimal model for predicting the relationship between the dependent and independent variables [IX]. A decision tree is a kind of classifier, also referred to as a graph framework, in which each leaf node of the tree is associated with a class label and the inside node indicates a feature [XVI].

E. Evaluation Metrics

The selection of the optimal algorithm for this particular dataset will be guided by an assessment of various performance metrics, encompassing AUC-ROC, accuracy, sensitivity, specificity, and confusion matrices. Equations 1-3 provide formulae for these metrics.

Confusion Matrix: A confusion matrix is used to analyze how well a model can distinguish between two classes. It shows how many times the model correctly predicted each class (positive, negative) and how often it made mistakes. Both the predicted and actual values are displayed in a two-row, two-column matrix.

$$Accuracy = \frac{TP+TN}{TN+FP+FN+TP} \quad (1)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (2)$$

$$Specificity = \frac{TN}{TN+FP} \quad (3)$$

IV. Results and Discussions

We have implemented feature selection and classification for our analysis using Python. We have found 10 important features using an extra tree classifier. The selected features are LEDD_CAT, mean_putamen, updrs3_score, mean_striatum, PC1, updrs2_score, age, mean_caudate and mutation_status in ascending order of their gain

score. Fig.1 depicts feature importance with their respective gain score. The selected features have been utilized to classify controls into PD-affected and HC using RF, LR, and DT classifiers. Each classifier has been trained using the trained data, and its performance has been predicted using the test data. There is a 70:30 ratio between train data and test data. Table 2 and Fig.2 demonstrate the accuracy, specificity, sensitivity, and AUC values of RF, LR, and DT classifiers. RF has shown the best performance concerning accuracy (96.12%), sensitivity (95.59%), and specificity (98.16%) while LR depicted the best AUC value of 98.33%, and accuracy, sensitivity, and specificity are 95.73%, 93.25% and 95.23% respectively. DT has demonstrated accuracy (93.91%), sensitivity (91.25%), and specificity (91.81%).

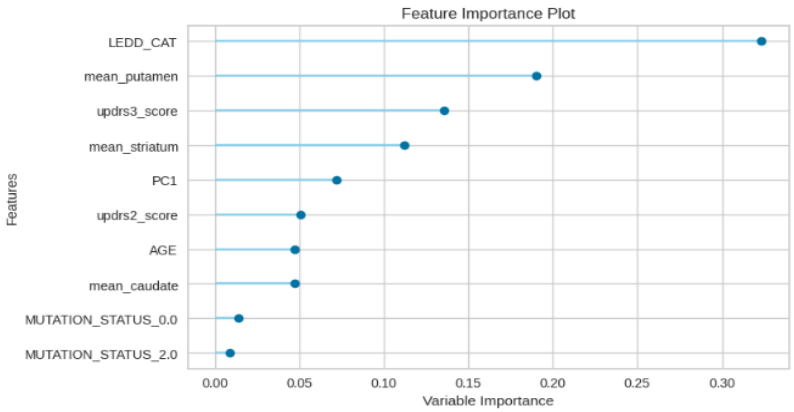

Fig. 1. Feature Importance Plot with gain score.

Table 2: Performance Metrics

Classifier	Accuracy %	Sensitivity %	Specificity %	AUC %
RF	96.12	95.59	95.34	98.16
LR	95.73	93.25	95.23	98.33
DT	93.91	91.25	91.81	93.65

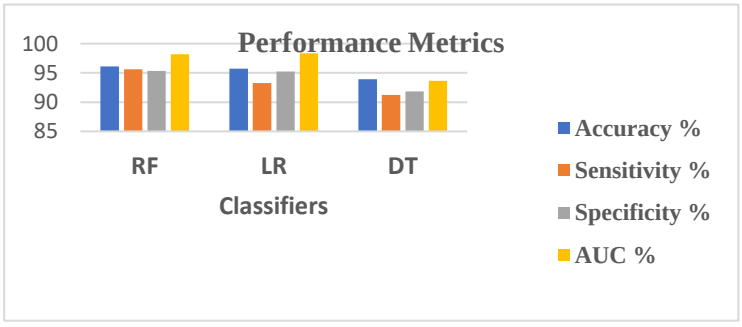

Fig. 2. Performance Metrics (Proposed Study).

Fig.3, Fig.4 and Fig.5 are demonstrating the ROC curves and confusion matrices of RF, LR, and DT respectively. RF, LR, and DT have depicted AUC values of 98.16%, 98.33% and 93.65% respectively. LR has shown the highest AUC value. The confusion matrix depicts that RF has made 316 accurate predictions out of 331 followed by LR with 312 correct predictions while DT has made the least accurate predictions of 303.

All of the classifiers have done well in PD classification, as seen in the figures below. Essentially, by eliminating redundant sets with little to no effect on performance, feature selection procedures help to conserve time and space. The study has demonstrated the reliability of an extra tree classifier as an efficient feature selection technique. RF has outperformed all three classifiers.

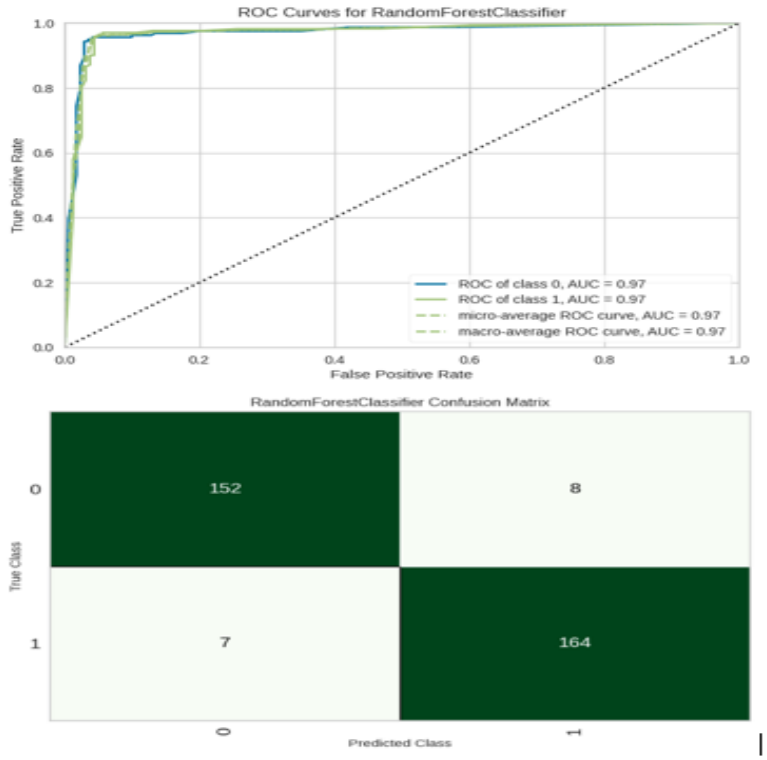

Fig. 3. The ROC and confusion matrix for Random Forest.

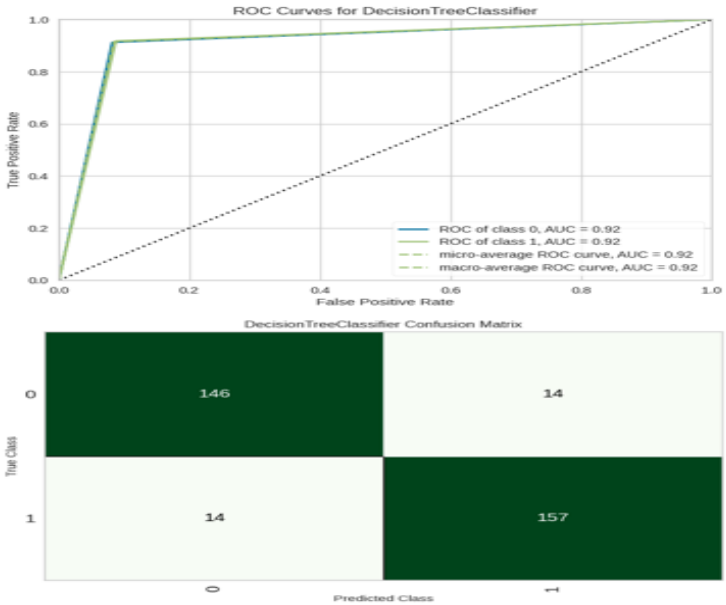

Fig. 4. The ROC and confusion matrix for Logistic Regression.

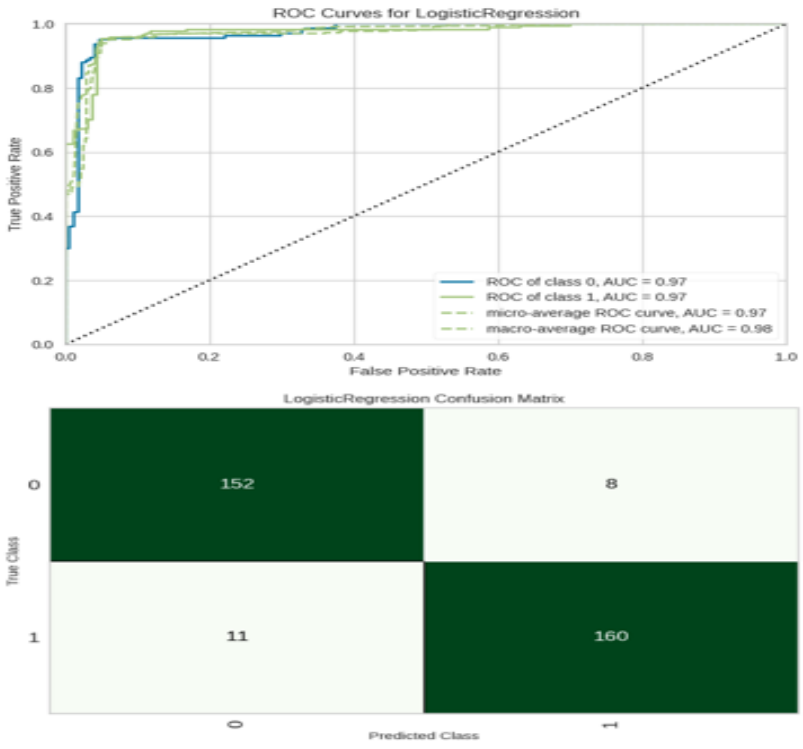

Fig. 5. The ROC and confusion matrix for the Decision Tree.

Table 3. demonstrates a comparison of the proposed study with prior research.

Table 3: Comparison of the proposed study with previous studies

Study	Modality Used	Number of Subjects	Feature Selection Method	Classifier Used	Accuracy %
Pantaleo et al. [XII]	Genes	550	Nested FS	RF, XGBoost	72
Oh et al. [XI]	EEG	40 (20 PD, 20 HC)	----	CNN	88.25
Rasheed et al. [XVII]	Speech	31(23 PD, 8 HC)	PCA	BPVAM (Back Propagation Algorithm with Variable Adaptive Momentum)	97.5
Sakar and Kursun [XX]	Speech	32 (24 PD, 8 HC)	Mutual information + maximum relevance–minimum redundancy	SVM	92.75
Abdhulay et al. [I]	Gait	166 (93 PD, 73 HC)	Peek detection, Pulse duration	Medium Tree, Medium Gaussian SVM	92.70
Yaman et al. [XXVI]	Speech	80 (40 PD, 40 HC)	ReliefF Algorithm	SVM, KNN	91.25
Proposed Study	**Heterogeneous Clinical Data**	**1103 (569 PD, 534 HC)**	**Extra Tree**	**RF, DT, LR**	**96.12**

The results indicate that utilizing clinical heterogeneous data has proved to be a successful biomarker for PD classification and it is novel as there are no other studies that have utilized heterogeneous clinical data for PD classification. Furthermore, the proposed study has included a larger number of subjects compared to studies from the literature. In addition, ETC as the feature selection and RF for the classification has rendered promising results.

V. Conclusion

This study has proposed ETC as a feature selection technique to classify controls into PD-affected and HC. The dataset used is heterogeneous containing clinical data, CSF-based proteomes analysis features, and gene mutation information. The disease_status feature in the dataset is binary, with 1 denoting PD and 0 denoting HC. There are 1103 instances in total with 20 features. The extra tree classifier has selected 10 features based on feature importance. Further selected features have been used for PD classification using RF, LR, and DT classifiers. RF has depicted the best performance with 96.12% accuracy. LR and DT have also shown promising results by achieving accuracy values of 95.73% and 93.91% respectively.

The future direction of the study could be further integrating genomics, and transcriptomics data and exploring different AutoML tools to diagnose PD. Different feature selection and optimization techniques could be applied further to enhance the study.

Conflict of Interest:

There was no relevant conflict of interest regarding this paper.

References

I. Abdulhay Enas, N. Arunkumar, Kumaravelu Narasimhan, Elamaran Vellaiappan, and V. Venkatraman. : 'Gait and tremor investigation using machine learning techniques for the diagnosis of Parkinson disease'. *Future Generation Computer Systems*. Vol. 83, pp. 366-373, (2018). 10.1016/j.future.2018.02.009

II. Bloem, Bastiaan R., Michael S. Okun, and Christine Klein. : 'Parkinson's disease'. *The Lancet*. Vol. 397, no. 10291, pp. 2284-2303, (2021). 10.1016/S0140-6736(21)00218-X

III. Bonifati Vincenzo. : 'Genetics of Parkinson's disease–state of the art, 2013'. *Parkinsonism & related disorders*. Vol. 20, pp. S23-S28, (2014). 10.1016/S1353-8020(13)70009-9

IV. Caramia Carlotta, Diego Torricelli, Maurizio Schmid, Adriana Muñoz-Gonzalez, Jose Gonzalez-Vargas, Francisco Grandas, and Jose L. Pons. : 'IMU-based classification of Parkinson's disease from gait: A sensitivity analysis on sensor location and feature selection'. *IEEE journal of biomedical and health informatics*. Vol. 22(6), pp. 1765-1774, (2018). 10.1109/JBHI.2018.2865218

V. Chapuis S., Ouchchane L., Metz O., Gerbaud L., & Durif F., : 'Impact of the motor complications of Parkinson's disease on the quality of life'. *Movement disorders: official journal of the Movement Disorder Society*. Vol. 20(2), pp. 224-230, (2005). 10.1002/mds.20279

VI. Dhiman Poonam, Vinay Kukreja, Poongodi Manoharan, Amandeep Kaur, M. M. Kamruzzaman, Imed Ben Dhaou, and Celestine Iwendi. : 'A novel deep learning model for detection of severity level of the disease in citrus fruits'. *Electronics*. Vol. 11(3), pp. 495, (2022), 10.3390/electronics11030495

VII. Kaiser Sergio, Luqing Zhang, Brit Mollenhauer, Jaison Jacob, Simonne Longerich, Jorge Del-Aguila, Jacob Marcus et al., : 'A proteogenomic view of Parkinson's disease causality and heterogeneity'. *npj Parkinson's Disease*. Vol. 9(1), pp. 24, (2023).10.1038/s41531-023-00461-9

VIII. Kaushal Chetna, and Anshu Singla. : 'Automated segmentation technique with self-driven post-processing for histopathological breast cancer images'. *CAAI Transactions on Intelligence Technology*. Vol. 5(4), pp. 294-300, (2020). 10.1049/trit.2019.0077

IX. Kurt Imran, Mevlut Ture, and A. Turhan Kurum. : 'Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease'. *Expert systems with applications*. Vol. 34(1), pp. 366-374, (2008). 10.1016/j.eswa.2006.09.004

X. Markello Ross D., Golia Shafiei, Christina Tremblay, Ronald B. Postuma, Alain Dagher, and Bratislav Misic. : 'Multimodal phenotypic axes of Parkinson's disease'. *npj Parkinson's Disease*. Vol. 7(1), pp. 6, (2021). 10.1038/s41531-020-00144-9

XI. Oh Shu Lih, Yuki Hagiwara, U. Raghavendra, Rajamanickam Yuvaraj, N. Arunkumar, M. Murugappan, and U. Rajendra Acharya. : 'A deep learning approach for Parkinson's disease diagnosis from EEG signals'. *Neural Computing and Applications*. Vol. 32, (2020). 10927-10933. 10.1007/s00521-018-3689-5

XII. Pantaleo Ester, Alfonso Monaco, Nicola Amoroso, Angela Lombardi, Loredana Bellantuono, Daniele Urso, Claudio Lo Giudice et al. : 'A machine learning approach to Parkinson's disease blood transcriptomics'. *Genes*. Vol. 13(5), pp. 727, (2022). 10.3390/genes13050727

XIII. Parmar Aakash, Rakesh Katariya, and Vatsal Patel. : 'A review on random forest: An ensemble classifier'. *International conference on intelligent data communication technologies and internet of things (ICICI)*. pp. 758-763, 2018. Springer International Publishing. 2019. 10.1007/978-3-030-03146-6_86

XIV. Patrician Patricia A., : 'Multiple imputation for missing data'. *Research in nursing & health*. Vol. 25(1), pp. 76-84, (2002). 10.1002/nur.10015

XV. Pol Urmila R., and Tejshree U. Sawant. : 'Automl: Building An Classfication Model With Pycaret'. *Ymer*. Vol. 20, pp. 547-552, (2021).

XVI. Priyanka and Dharmender Kumar. : 'Decision tree classifier: a detailed survey'. *International Journal of Information and Decision Sciences*. Vol. 12(3), pp. 246-269, (2020). 10.1504/IJIDS.2020.108141

XVII. Rasheed Jawad, Alaa Ali Hameed, Naim Ajlouni, Akhtar Jamil, Adem Özyavaş, and Zeynep Orman. : 'Application of adaptive back-propagation neural networks for Parkinson's disease prediction'. *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, IEEE, pp. 1-5. 2020. 10.1109/ICDABI51230.2020.9325709

XVIII. Sabherwal G., Kaur A., : 'Machine learning and deep learning approach to Parkinson's disease detection: present state-of-the-art and a bibliometric review'. *Multimedia Tools and Applications*. (2024). 10.1007/s11042-024-18398-3

XIX. Sachdeva Ravi Kumar, Tushar Garg, Gagandeep Singh Khaira, Dikshant Mitrav, and Rakesh Ahuja. : 'A Systematic Method for Lung Cancer Classification'. *10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), 2022*. IEEE, 2022. pp. 1-5. 10.1109/ICRITO56286.2022.9 964778

XX. Sakar C. Okan, and Olcay Kursun. 'Telediagnosis of Parkinson's disease using measurements of dysphonia'. *Journal of medical systems*. Vol. 34 pp. 591-599, (2010). 10.1007/s10916-009-9272-y

XXI. Schapira Anthony HV, K. Ray Chaudhuri, and Peter Jenner. : 'Non-motor features of Parkinson disease'. *Nature Reviews Neuroscience*. Vol. 18(7), pp. 435-450, (2017). 10.1038/nrn.2017.62

XXII. Sharaff Aakanksha, and Harshil Gupta. : 'Extra-tree classifier with metaheuristics approach for email classification'. *Advances in Computer Communication and Computational Sciences: Proceedings of IC4S 2018*. Springer Singapore, 2019. pp. 189-197. 10.1007/978-981-13-6861-5_17

XXIII. Shetty Sachin, and Y. S. Rao. : 'SVM based machine learning approach to identify Parkinson's disease using gait analysis'. *2016 International conference on inventive computation technologies (ICICT)*. IEEE, vol. 2, pp. 1-5. 2016. 10.1109/INVENTIVE.2016.7824836

XXIV. Sihombing Denny Jean Cross, Jawangi Unedo Dexius, Jonson Manurung, Mendarissan Aritonang, and Harni Seven Adinata. : 'Design and Analysis of Automated Machine Learning (AutoML) in PowerBI Application Using PyCaret'. *2022 International Conference of Science and Information Technology in Smart Administration (ICSINTESA)*. IEEE. pp. 89-94. 2022. 10.1109/ICSINTESA56431.2022.10041543

XXV. Tsukita Kazuto, Haruhi Sakamaki-Tsukita, Sergio Kaiser, Luqing Zhang, Mirko Messa, Pablo Serrano-Fernandez, and Ryosuke Takahashi. : 'High-throughput CSF proteomics and machine learning to identify proteomic signatures for Parkinson disease development and progression'. *Neurology*. Vol. 101(14), pp. e1434-e1447, (2023). 10.1212/WNL0000000000207725

XXVI. Yaman Orhan, Fatih Ertam, and Turker Tuncer. : 'Automated Parkinson's disease recognition based on statistical pooling method using acoustic features'. *Medical Hypotheses*. Vol. 135, 109483 (2020). 10.1016/j.mehy.2019. 109483

XXVII. Zhu Biqing, Dominic Yin, Hongyu Zhao, and Le Zhang. : 'The immunology of Parkinson's disease'. *Seminars in Immunopathology*, vol. 44(5), pp. 659-672. Berlin/Heidelberg: Springer Berlin Heidelberg, 2022. 10.1007/s00281-022-00947-3