Predicting Parkinson’s Disease using Machine Learning and Ensemble Learning Algorithms

Vibha B Hegde
Dept. of Data Science & Computer Applications
Manipal Institute of Technology
Manipal Academy of Higher Education
Manipal, Karnataka, India-576104
vibha.mitmpl2023@learner.manipal.edu

***Abstract*—Parkinson’s disease is a neurological condition that impairs movements and causes shaking, rigidity, difficulty in balancing and coordination, and problems with speech. This gets worse over time. The most prevalent symptoms that can be identified by looking at patients data is variations in the patient’s voice. Parkinson’s disease is an important study topic since early diagnosis can lead to better and improved health. This paper illustrate a few approaches based on Machine learning, which can be applicable in predicting Parkinson’s disease. Additionally, few ensemble approaches are considered for predictive disease analytics. Normally, an Ensemble approach makes predictions that are more accurate than those of a single model by combining multiple models. The generated results can be compared with those of other machine learning models to obtain the most optimized and efficient model to estimate the severity of Parkinson’s disease.**

***Index Terms*—Parkinson’s disease, machine learning, ensemble learning, voice dataset**

I. Introduction

People across the globe suffer from Parkinson’s disease(PD). Those who suffer from PD or who are at potentiall risk of getting the condition would benefit from a suitable, quick, and efficient way to assess the existence and extremity of the disease’s symptoms. There are two main categories of PD symptoms: Movement related referred to as motor symptoms and unrelated to movement called as non motor symptoms. Lack of movement (bradykinesia), rigidity, improper balance and coordination and voice dysfunction comes under motor symptoms. Non motor symptoms include depression, sleep difficulties, lack of odor, cognitive decline. Voice measures can be effective technique for PD detection since they may be used to distinguish between those who having the disease and those who are not.

Till date, the reason behind PD is not known. It is noteworthy that prompt and early recognition of PD promotes rapid treatment and alleviate symptoms significantly [1].

A hybrid machine learning model is more stable and performs better compared to normal machine learning model, in terms of producing accurate results, minimizing noise in data, etc. [2].

Machine Learning allows us to build algorithms that automatically identify patterns in the available data without explicit instructions. Machine learning is sometimes confused with artificial intelligence, mainly due to its ability for learning and decision-making, but in reality, it is a subset of artificial intelligence. Up to the late 1970s, it was a part of Artificial Intelligence’s evolution. It then began to change and evolve on its own. [3], [4]. Several technologies can be used to early detect PD. Because deep learning and machine learning-based methods can manage large amounts of data and produce results with high accuracy, they are utilized in the diagnosing of PD [5]. Machine learning depends entirely on data. The more data there is, the more effective machine learning is. So, using this data computer systems discover correlations hidden within it. This ability can be extended through programming and developing intelligent machine Learning algorithms [6]. Machine learning is helpful in extracting and analyzing relevant information regarding PD because of its data-driven methodologies. Additionally, machine learning methods are crucial for diagnosing and treating PD and expedite the decision-making process. Finding relevant data is crucial in machine learning concept. The result will be inaccurate if a reliable data source is not provided. Additionally, data quality is also very important [7].

To ensure early PD diagnosis, vocal measurements could be used. Depending on the vocal parameters of the impacted person, a thorough analysis can be carried out. These features can be fed to the ensemble learning model and to other models like Decision tree, KNN, Random forest, SVM, etc., to gauge seriousness of PD.

The outline of the paper is: Literature review is covered in Section II. The PD dataset and Methodology has been discussed in Section III. Section IV and V discusses the PD prediction results and conclusion respectively.

II. Literature Review

There have been numerous studies conducted to predict PD. To determine PD, a variety of machine learning models are employed.

Váradi, C. [8] has proposed a review with an emphasis on nonmotor symptoms, this article discusses the development of important aspects in the diagnosis of PD. These symptoms are common in PD. They play a part in the patients change in emotional and physical status because of weariness, excessive diaphoresis, hypersalivation, voice tone changes, etc. Symptoms of PD include bradykinesia, rest tremor, rigidity, and loss of postural reflexes. These primary motor traits are also used to identify patients in modern clinical practice. The effective use of levodopa medication has shown that the nonmotor characteristics are the primary causes of patient weakness in PD, and that they may arise before motor symptoms appear, during disease progression. The most commonly used clinical standards are based on motor symptoms alone, despite the fact that effective use of dopaminergic treatment has highlighted the significance of nonmotor features. The guidelines are precise and comprehensive, but a primary drawback is the requirement for highly qualified The group of patients is comparatively older, and often has many diseases, which may complicate using stringent clinical diagnostic standards. More data is needed to accurately identify diseases.

Carpi et al. [9] has proposed a study to examine the existence of Non motor symptoms (NMS) and Non-motor fluctuations (NMF) in PD patients employing the Non-Motor Fluctuation Assessment questionnaire (NoMoFa), which was recently validated, and to evaluate their relationships with disease attributes and motor impairment. According to this observational study which pointed out that in a sample of PD patients receiving pharmacological treatment, non-motor fluctuations were noted by up to 32% of the patients. These observations verify that the NMS and its variations are significant factors to consider in PD supervision and imply that certain pharmacological and non-pharmacological approaches may be needed for their treatment. To gather systematic evidence regarding NMS and NMF, Future studies need to make use of specialized procedures like the NoMoFa. Huge sample sizes and designs are needed for additional research to validate the findings, consistency and applicability to clinical practice.

Ensemble is a machine learning method that produces more accurate predictions combining multiple models and improves the results as compared to a single model. The study by Mahajan.P et al. [10] aims to help academics an improved ensemble model for analytics related to disease prediction and to better understand current developments of models for predicting disease that incorporate this method. The accuracy of the underlying classifier is increased when additional machine learning algorithms are applied using an ensemble technique.

Several ensemble techniques are used by Nazmun et al. [11] to detect liver disease. It primarily focuses on evaluating and diffrentiating the effectiveness of various ensemble techniques like AdaBoost, LogitBoost, BeggRep, BeggJ48 and Random Forest.

Latha et al. [12] used ensemble classification, by combining multiple classifiers. Heart disease information was used. This studies the difference in performance of ensemble techniques, in terms of accuracy. Adam et al. [13] examines how nanotechnology platforms like nanobiosensors and nanomedicine, have enhanced diagnosis and treatment of PD. The affordability, portability, and rapid and accurate analysis are their key features. In addition, nanoparticles can be transported via nanotechnology. There is limited application of nanobiosensors in biological systems because of their unique technology.

PD is still diagnosed clinically, based on the physician's capacity to identify its typical symptoms, especially in the starting stages as mentioned in the work by Pang et al. [14]. PD is assumed to be caused by an intricate interaction between environmental and inherited variables, but its exact cause is yet unknown.

PD is a complex and diverse set of condition, influenced by a patient's aging, genetics, and exposure to the environment in different ways. An individual's genetic vulnerability to pathogenic variations and common risk variants, which affect resistance to environmental variables and the aging process, will determine their risk of PD. Finding possible ways for disease modification may be easier with an understanding of the processes and events that cause initiation and progress of PD.

Raccagni et al. [15] proposed a review of postural dysfunction, gait, and balance in PD, concentrating on aspects of differential diagnosis. Axial signs such not being able to ride a bike, or frequent falls should warn the neurologist and result in a distinctive diagnosis of non-PD conditions. In order to distinguish atypical parkinsonian diseases (APD) from PD, the timed up and go test (TUG-test), tandem gait test, and retropulsion test could be used. Further research is required in order to verify this test battery.

The Covid virus infection has caused people and global medical system remarkably. Rai et al. [16] has mentioned in the study that more research is required to illustrate the comprehensive Human coronavirus (HCoV) mechanism of infection that results in damage of neurons, especially if HCoV is indeed able to take advantage of the related CNS invasion routes. If so, it could be intriguing to examine how certain processes interact with the particular symptoms of neurological conditions. This hasn't been resolved yet.

Selvaraj et al. [17] proposed a review which examines the molecular genetics of PD. The cause of PD is the presence of neuronal intracellular. Genetic researches in PD gobally are anticipated to yield a novel medication for the long-term recovery.

Kodali et al. [18] proposed a study where a systematic comparison is conducted, where Speech-based PD severity classification was studied. This study has solved issues with multi-class PD severity level classification. Larger databases is required to create a clinically useful automatic severity level classification system.

Indu et al. [19] utilized traditional kNN to identify PD with Gait, Handwriting, and Voice parameters. This method replaces voting and neighboring points with the concepts of weights and -neighborhood to estimate the test's unknown class samples. Regardless of sample sizes, the updated kNN algorithm is more effective in identifying Parkinson's patients than other supervised classifiers. The work offers a productive alteration to the conventional kNN for precise PD diagnosis. However, the work is susceptible to a lot of data overlap and the suggested kNN algorithm has an O(n2) time complexity.

Haq et al. [20] used a system driven by machine learning using predictive algorithm called SVM was employed to forecast PD. The features L1-norm SVM selection was employed for relevant and choosing closely linked features to accurately classify those with PD and healthy. The suggested method closes a gap in feature selection and classification., utilizing data from audio recordings by appropriately aligning with the experimental framework.

Hamzehei et al. [21] utilized ML techniques and cloud. Four machine learning techniques is applied to data on PD to evaluate the techniques, performance, and pinpoint the key elements that might be utilized in the prediction. According to the estimated weight of the algorithm, the most important factor predicting Total UPDRS is Motor UPDRS. The idea of the cloud is thought to improve accessibility and accuracy while reducing time complexity. A neural network on the cloud and a variety of PD datasets can be employed for future expansion, and finally a review of the approaches.

Bansal et al. [22] used handwritten data where the diagnosis was made using a categorization technique that was age- and gender-dependent utilizing online handwritten samples recorded from individuals with PD and healthy controls. A SVM ranking algorithm is used to showcase unique characteristics to their dominance in gender and age range for Parkinson's diagnosis. Combining the age and gender information showed promising results in classification.

Chen et al. [23] used a machine-learning framework. Using smartphone sensors, signals were extracted from patients performing numerous active testing and passive monitoring , such as voice, dexterity (ability to complete tasks with hands), tremor, gait (walking pattern), and imbalance. The Machine-Learning framework includes an elastic-net regularization-based generic model and a two-step feature selection process which were categorized using a framework for automated disease evaluation.. The limitation is that only behavioral features that are linearly related are captured. To identify the non-linear feature architecture associated with PD, future research employing non-linear parametric approaches could be helpful.

Gomathy et al. [24] suggested a work with an objective to show how accurate early detection of PD identification is. This work uses several algorithms like XGBoost, Decision Tree Classifier and Navie Bayes to test the ability of motor function of the patient with PD. stages.

Quan et al. [25] put forth a work where the dynamic and static elements of speech are considered for PD detection and a deep learning-based approach is suggested. The experimental findings demonstrated that, in comparison to other models, this increases the accuracy of PD identification. As future work, it is suggested that this method can be applied for PD stage classification to investigate its suitability for multi-label classification.

III. Methodology

As of now, there are no readily available blood or laboratory tests to detect PD in non-hereditary patients. However, doctors usually diagnose the disease by various symptoms and diagnostic tests. PD is an important topic since early diagnosis can lead to better health of patients. The use of speech data from subjects will benefit in the development of non-invasive diagnostics.

Fig 1 depicts the general structure of the method for detecting PD. The dataset includes several voice measurements of people made during a six-month period utilizing a telemonitoring device that remotely keeps an eye on people's symptoms and indicators. The flowchart illustrates the proposed process from data gathering to model selection. Data may be unordered and might contain many missing values, duplicates, noisy data and outliers. Data Pre- processing should be done before further using the data.

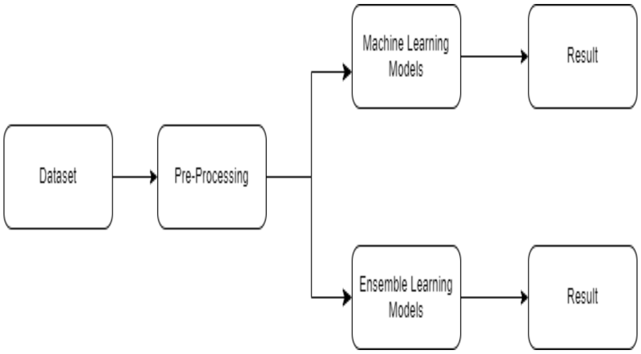

Fig. 1. Block Diagram

Ensemble learning is a machine learning methodology that groups many individual models also referred to as base learners or base models to improve the overall execution and generalization of a predictive model. The ensemble model is typically more robust and accurate than any single model in the ensemble. The machine learning and ensemble models are compared and results are computed. The proposed method from gathering data to choosing a model is depicted in Fig 2

A. Collection of data

The dataset consists of a varying biomedical vocal measurements collected from 42 people with PD in early-stages who participated in a half-year trial of a telemonitoring tool for tracking the course of symptoms remotely. The recordings were made in the patients homes. The two main target variables are - Motor UPDRS (Unified Parkinson's Disease Rating Scale) and Total UPDRS - both assessment scores utilized in assessing the seriousness and progression of PD. The Motor UPDRS is primarily focused on motor function and comprises of various aspects of movement. The Total UPDRS is an all-inclusive rating scale that covers both motor and non-motor symptoms connected with PD.

There are total 5,875 voice recordings collected from these people. Predicting the motor and overall UPDRS scores is the primary goal of the data. The information is in CSV ASCII format. Every row in the CSV file specifies one voice measurement. There are approximately 200 recordings for each patient and the his/her subject number is listed in the first column. The

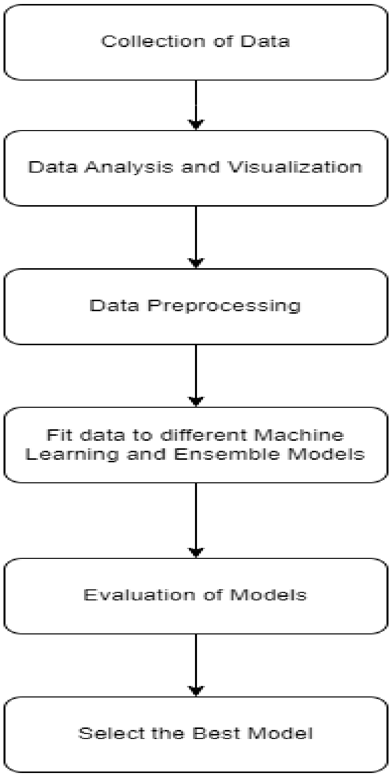
Collection of Data
Data Analysis and Visualization
Data Preprocessing
Fit data to different Machine Learning and Ensemble Models
Evaluation of Models
Select the Best Model

Fig. 2. Flow Chart

dataset is divided into two. One with motor_UPDRS and other with total_UPDRS, including all other voice measurements. And results are calculated separately for motor_UPDRS and total_UPDRS.

The attributes in dataset are:
subject# - a unique integer representing each subject
1. age - Persons age
2. sex - Persons gender ’0’ is male, ’1’ is female
3. test_time - Duration since enrollment in the experiment. The number of days since recruiting is the integer portion.
4. motor_UPDRS - The motor UPDRS score of the person, interpolated linearly
5. total_UPDRS - The total UPDRS score of the person, interpolated linearly
6. Jitter(%), Jitter(Abs), Jitter - RAP, Jitter - PPQ5, Jitter - DDP : Numerous metrics for fundamental frequency variation Shimmer, Shimmer(dB), Shimmer - APQ3, Shimmer - APQ5, Shimmer - APQ11, Shimmer - DDA : Multiple amplitude variation measurements
7. NHR,HNR - Two measurements of the voice’s , noise to tonal component ratio
8. RPDE - A measure of nonlinear dynamical complexity
9. DFA - Exponent of signal fractal scaling
10. PPE - A nonlinear indicator of the underlying frequency fluctuation

B. Data Analysis and Visualization

Several techniques are used in data analysis to search for patterns, clusters, or other relationships between different types of data. It is critical to understand data. Data visualization is putting data into a chart, graph for analysis and interpretation. Python libraries like Matplotlib, Seaborn etc. can be used for Data visualization.

C. Data Preprocessing

The data that is collected cannot be directly used for analysis because the data may be unordered and might contain many missing values, duplicates, noisy data and outliers. The problems like missing values, duplicate removal, outliers,etc. if present must be either removed or treated accordingly. Data Preprocessing step includes removing useless and incomplete data.

D. Fit Data to different Machine Learning and Ensemble learning Models

Data must be fit to different machine and ensemble learning models for PD prediction. Model fit indicates how well a learning model generalizes to the data it was trained on.

E. Evaluation of Models

Model evaluation is done to determine how well the model performs its task. The dataset after cleaning, must be split into train and test data or a complex stratified k-fold strategy. In case of classification problems, a model’s performance can be evaluated using accuracy, confusion matrix which is a tabular summary of the prediction. And for regression problems, metrics like mean square error (mse), root mean square error (rmse), etc; can be used.

F. Select the best model

Model selection is the process of deciding which algorithm and model architecture works best with a certain data set in hand. Choosing a model depends on evaluation results. Model selection is an important step in machine learning since different models have varying complexities, underlying assumptions, and characteristics.

IV. Experimentation and Results

42 people with early-stage PD were included in the entire dataset to predict PD. They took part in a trial lasting six months of a telemonitor-ing tool for remote symptom progression monitoring. There are total 5,875 voice measurements. Predicting the motor and overall UPDRS scores is the primary goal of the data.

Machine Learning and Ensemble Learning methods were used to predict the scores. Machine learning models : Linear Regression, SVM Regressor, kNN Regressor, Decision Tree Regressor were implemented. Ensemble Learning models - Random Forest Regressor, AdaBoost, Gradient Boosting, Bagging Regressor were implemented. Evaluation metrics such as Root mean square error.

(Rmse) and R-Squared (R2 score) are employed to assess the models' effectiveness. Both the Motor and Total UPDRS scores are assessed independently for each model. The corresponding tables below display the results.

TABLE I
RMSE and R2 Score for Motor UPDRS

ML Models	RMSE Score	Rˆ2 Score
Linear Regression	7.484263	0.122437
Support Vector Machine Regressor	7.709999	0.068701
K-Nearest Neighbors Regressor	5.791841	0.474450
Decision Tree Regressor	2.143185	0.928039

TABLE II
RMSE and R2 Score for Total UPDRS

ML Models	RMSE Score	Rˆ2 Score
Linear Regression	9.659540	0.157981
Support Vector Machine Regressor	9.807376	0.132187
K-Nearest Neighbors Regressor	7.346752	0.512921
Decision Tree Regressor	2.405221	0.947794

TABLE III
RMSE and R2 Score for Motor UPDRS

Ensemble Models	RMSE Score	Rˆ2 Score
Random Forest Regressor	1.317416	0.972809
Gradient Boosting Regressor	3.958634	0.754489
AdaBoost Regressor	6.269520	0.384187
Bagging Regressor	1.567930	0.961485

TABLE IV
RMSE and R2 Score for Total UPDRS

Ensemble Models	RMSE Score	Rˆ2 Score
Random Forest Regressor	1.613826	0.976497
Gradient Boosting Regressor	4.920833	0.781482
AdaBoost Regressor	8.333343	0.373317
Bagging Regressor	1.795582	0.970905

V. Conclusion

Machine learning and other technologies are being developed and improved to predict PD. Ensemble learning is proved to be a more effective method. The Machine learning models are assessed using Root mean square error (Rmse) and R-Squared (R2 score) as shown in Table I and II. The results of Ensemble learning models are shown in Table III and IV. Ensemeble Models - Random Forest Regressor and Bagging Regressor are better models compared to other Machine learning models, as it has the lowest RMSE and highest R2 Score. Random Forest Regressor has RMSE of 1.6 and an R2 score of 0.97. Hence proven to be a better model.

Ensemble learning approach might work well for some specific data like voice measurements, etc. But it might not be able to provide accurate results. Other datasets such as Medical image datasets might be utilized to determine PD effectively. As a future scope, combining several data paradigms (e.g.imaging, genetic) and technologies can offer a thorough evaluation of the risk, detection and prognosis of PD.

References

[1] W. Wang, J. Lee, F. Harrou and Y. Sun, ”Early Detection of Parkinson’s Disease Using Deep Learning and Machine Learning,” in IEEE Access, vol. 8, pp. 147635-147646, 2020, DOI: 10.1109/ACCESS.2020.3016062.

[2] Saeid Sheikhi, Mohammad Taghi Kheirabadi, ”An Efficient Rotation Forest-Based Ensemble Approach for Predicting Severity of Parkinson’s Disease”, Journal of Healthcare Engineering, vol. 2022, Article ID 5524852, 9 pages, 2022. https://doi.org/10.1155/2022/5524852

[3] Janiesch, C., Zschech, P. Heinrich, K. Machine learning and deep learning. Electron Markets 31, 685–695 (2021). https://doi.org/10.1007/s12525-021-00475-2

[4] Zhang, Y., Ling, C. A strategy to apply machine learning to small datasets in materials science. npj Comput Mater 4, 25 (2018). https://doi.org/10.1038/s41524-018-0081-z

[5] Gandomi, A.H.; Chen, F.; Abualigah, L. ”Machine Learning Technologies for Big Data Analytics.Electronics2022,11 421. https://doi.org/10.3390/electronics11030421

[6] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x

[7] Mei J, Desrosiers C, Frasnelli J. Machine Learning for the Diagnosis of Parkinson’s Disease: A Review of Literature. Front Aging Neurosci. 2021 May 6;13:633752. doi: 10.3389/fnagi.2021.633752. PMID: 34025389; PMCID: PMC8134676.

[8] Váradi, C. Clinical Features of Parkinson’s Disease: The Evolution of Critical Symptoms. Biology 2020, 9, 103.

[9] J. Carpi, M., Pierantozzi, M. Cofano, S. Fernandes, M. Cerroni, R. De Cillis, F. Mercuri, N.B. Stefani, A. Liguori, C. Both Motor and Non-Motor Fluctuations Matter in the Clinical Management of Patients with Parkinson’s Disease: An Exploratory Study. J. Pers. Med. 2023, 13, 242.

[10] Mahajan, Palak, Shahadat Uddin, Farshid Hajati, and Mohammad Ali Moni. 2023. ”Ensemble Learning for Disease Prediction: A Review” Healthcare 11, no. 12: 1808. https://doi.org/10.3390/healthcare11121808

[11] Nahar, N.; Ara, F.; Neloy, M.A.I.; Barua, V.; Hossain, M.S.; Andersson, K. A comparative analysis of the ensemble method for liver disease prediction. In Proceedings of the 2019 2nd International Conference on Innovation in Engineering and Technology (ICIET), Dhaka, Bangladesh, 23–24 December 2019; pp. 1–6.

[12] Latha, C.B.C.; Jeeva, S.C. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Inform. Med. Unlocked 2019, 16, 100203.

[13] Adam, H., Gopinath, S.C.B., Md Arshad, M.K. et al. An update on pathogenesis and clinical scenario for Parkinson’s disease: diagnosis and treatment. 3 Biotech 13, 142 (2023).

[14] Pang, S.YY., Ho, P.WL., Liu, HF. et al. The interplay of aging, genetics and environmental factors in the pathogenesis of Parkinson’s disease. Transl Neurodegener 8, 23 (2019).

[15] Raccagni, C., Nonnekes, J., Bloem, B.R. et al. Gait and postural disorders in parkinsonism: a clinical approach. J Neurol 267, 3169–3176 (2020).

[16] Sachchida Nand Rai, Neeraj Tiwari, Payal Singh, Anurag Kumar Singh, Divya Mishra, Mohd. Imran, Snigdha Singh, Etrat Hooshmandi, Emanuel Vamanu, Santosh K. Singh, Mohan P. Singh, ”Exploring the Paradox of COVID-19 in Neurological Complications with Emphasis on Parkinson’s and Alzheimer’s Disease”, Oxidative Medicine and Cellular Longevity, vol. 2022, Article ID 3012778, 16 pages, 2022.

[17] Suganya Selvaraj, Shanmughavel Piramanayagam, “Impact of gene mutation in the development of Parkinson’s disease”, Genes Diseases, Volume 6, Issue 2, 2019, Pages 120-128, ISSN 2352-3042.

[18] Kodali, M., Kadiri, S., Alku, P. (2023). ”Automatic classification of the severity level of Parkinson’s disease: A comparison of speaking tasks, features, and classifiers”. Computer Speech and Language, 83, [101548].

[19] Indu, R., Dimri, S.C. Malik, P. ”A modified kNN algorithm to detect Parkinson’s disease. Network Modeling Analysis Health Informatics and Bioinformatics” 12, 24 (2023).

[20] A. U. Haq et al., ”Feature Selection Based on L1-Norm Support Vector Machine and Effective Recognition System for Parkinson’s Disease Using Voice Recordings,” in IEEE Access, vol. 7, pp. 37718-37734, 2019.

[21] Hamzehei S., Akbarzadeh O., Attar H., Rezaee K., Fasihihour N., Khosravi M.R. Predicting the total Unified Parkinson’s Disease Rating Scale (UPDRS) based on ML techniques and cloud-based update (2023).

[22] Gupta U, Bansal H, Joshi D. ”An improved sex-specific and age-dependent classification model for Parkinson’s diagnosis using handwriting measurement” Compute Methods Programs Biomed. 2020 Jun;189:105305.

[23] O. Y. Chén et al., ”Building a Machine-Learning Framework to Remotely Assess Parkinson’s Disease Using Smartphones,” in IEEE Transactions on Biomedical Engineering, vol. 67, no. 12, pp. 3491-3500, Dec. 2020.

[24] C K Gomathy, B. Dheeraj kumar Reddy, B. Varsha and B. Varshani, ”The Parkinson’s Disease Detection Using Machine Learning Techniques”, International Research Journal of Engineering and Technology (IRJET), vol. 08, no. 10, Oct 2021.

[25] C. Quan, K. Ren and Z. Luo, ”A Deep Learning Based Method for Parkinson’s Disease Detection Using Dynamic Features of Speech,” in IEEE Access, vol. 9, pp. 10239-10252, 2021, doi: 10.1109/ACCESS.2021.3051432.