International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 2, Mar-Apr 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Parkinson's Disease Prediction Using Machine Learning

Omkar Singh[1], Santosh Singh[2], Anjali Rasal[3]
1HOD(Department of Data Science),
2HOD(Department of Information Technology),3PG Students
1,3 Department of Data Science, 2Department of Information Technology,
Thakur College of Science and Commerce
Thakur Village, Kandivali(East), Mumbai-4000101, Maharashtra, India.
omkarsingh@tcsc.edu.in, sksingh@tcsc.edu.in, rasalanjali0205@gmail.com

Abstract:
Parkinson's disease (PD) is a progressive neurodegenerative disorder that affects millions globally, characterized by motor symptoms such as tremors, bradykinesia, and rigidity, as well as non-motor symptoms. Early diagnosis of PD is challenging due to its symptom overlap with other disorders, often resulting in delayed intervention. Recent deep learning (DL) advances offer promising approaches for enhancing early detection accuracy, leveraging large-scale datasets and advanced neural networks. This paper explores the application of DL techniques, including Random Forest(RF) and support vector Regression(SVR) to analyze clinical data, gait patterns, and medical imaging for PD diagnosis. We focus on supervised learning models trained on various multimodal data sources to improve diagnostic precision and identify potential biomarkers for early PD detection. Experimental results demonstrate the capability of DL models to achieve high classification accuracy, contributing to reliable and accessible diagnostic tools. Future work includes integrating additional data sources, refining algorithms to minimize computational costs, and implementing explainable AI for clearer clinical interpretations. Our findings underscore the transformative potential of deep learning in advancing Parkinson's disease diagnostics, supporting timely and personalized medical intervention. Current research aims to address the challenges posed by late-stage diagnosis and limited treatment options by focusing on biomarker discovery, genetic and molecular analysis, drug development, and computational modeling. Biomarker research seeks to identify indicators in blood, cerebrospinal fluid, and neuroimaging data that can enable early PD diagnosis before severe symptoms emerge. Genetic and molecular studies are advancing the understanding of hereditary and biochemical pathways involved in PD, informing personalized treatment strategies. Drug development efforts are exploring neuroprotective therapies to preserve or restore affected brain cells. Additionally, machine learning approaches are being utilized to analyze complex datasets, identify patterns in disease progression, and customize patient care. Environmental and lifestyle studies also play a role in examining external factors that may influence PD risk. Together, these approaches hold promise for transforming PD from a progressively debilitating condition into a manageable or preventable one, ultimately improving patient outcomes and reducing the societal impact of the disease.

Keywords: Treatment process, Life, EDA, RF, SVR treatment, model, motor symptoms, non-motor symptoms, PD, DL, Learning, approaches, early, disease, diagnosis, data, Al, Symptoms, identity, treatment, studies

1. Introduction:

Parkinson's disease (PD) is a brain condition that causes movement, mental health, sleep, pain, and other health issues. The exact cause of PD remains unknown, but a combination of genetic and environmental factors is thought to contribute. Small genetic mutations, such as those in the LRRK2 or SNCA genes, are linked to a small percentage of cases. At the same time, exposure to certain toxins like pesticides and heavy metals has been associated with an increased risk of developing PD. Current treatment options

focus on managing symptoms rather than curing the disease. Medications like Levodopa and deep brain stimulation (DBS) surgery help alleviate motor symptoms, while physical, speech and occupational therapies support functional abilities.

Other research aims to develop disease-modifying therapies that could slow or halt PD's progression. Scientists are exploring gene and stem cell therapies to restore dopamine function, as well as identifying biomarkers for earlier diagnosis. These efforts, alongside the search for neuroprotective drugs, represent promising steps toward understanding and potentially changing the course of PD. Parkinson's disease usually occurs in older people, but younger people can also be affected. Men are affected more often than women. The cause of PD is unknown, but people with a family history of the disease are at a higher risk. Exposure to air pollution, pesticides, and solvents may also increase risk. Informal carers spend many hours daily providing care for people living with PD, which can be overwhelming.

The motivation for researching Parkinson's disease (PD) is driven by the profound need to improve diagnosis, treatment, and quality of life for millions of individuals affected by this progressive and debilitating neurodegenerative disorder. With PD's prevalence increasing globally due to aging populations, the healthcare and societal burden of the disease continues to grow, creating urgency to find effective interventions.

2. Literature Review:

Parkinson's disease (PD) can be diagnosed using neuropathologic and histopathologic criteria. Classification can be done through literature review and selection based on sensitivity and specificity of clinical features. Clinical-pathologic studies are needed to investigate patients' frequency of occurrence, characteristics, and risk factors. Classifiers like neural networks, DMneural, regression, and decision trees are used for reliable diagnosis. Telemonitoring of PD using voice measurement is crucial for early diagnosis. Classification accuracy, Kappa Error, and Area under the Receiver Operating Characteristic Curve are used to assess PD relations' relevance and statistical significance to attributes[1]. Parkinson's disease is a neurodegenerative disorder affecting millions worldwide, causing symptoms like muscle fatigue, tremors, and dementia. Big Data (BD) is used to analyze this data, which includes properties like velocity, veracity, length, meaning, and variety. The data is heterogeneous, multi-source, and unreliable. Treatment judgments are crucial for diagnosis and computational performance. Parkinson's disease is a chronic brain condition with non-motor symptoms and motor symptoms. Movement conditions, such as depression and schizophrenia, are common causes. Data is stored and summarized using Hadoop-bigdata, utilizing high data analysis technologies[2]. Parkinson's disease (PD) is a significant degenerative disease affecting millions of seniors worldwide. Symptoms can vary from person to person and can be motor or non-motor. Non-motor symptoms include depression, sleep disorders, loss of smell, and cognitive impairment. PD complications are the 14th leading cause of death in the US. The economic burden of PD, including treatment, social security payments, and lost income, is estimated to be around $52 billion per year. Early detection of PD can facilitate rapid treatment and potentially lead to access to disease-modifying therapy[3].

Neurodegenerative diseases like Alzheimer's, Parkinson's, and Arthritic disease are prevalent, with Parkinson's disease being the second most common. Medical information is crucial for diagnosis, patient care, and research. Clinical decision intelligence aims to streamline data management across clinical practice, nursing, healthcare management, and administration. Machine learning-based methods are used for knowledge acquisition and evidence-based research, analyzing information from various sources and quality scores[4]. Parkinson's disease, a neurological disorder, causes nerve cell death in the substantia nigra, affecting men's voices and mating. It is second to Alzheimer's disease in neurological disorders. Expansion is expected, necessitating further research to identify treatments and develop appropriate testing frameworks[5]. Recent image generation and enhancement advancements have utilized in-depth learning techniques, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNNs are commonly used for image recognition and classification, while RNNs are recommended for sequential data representation. Deep learning algorithms have been applied to MRI imaging, making it more accessible from various perspectives. CNNs process images from the

base, detecting small features and then searching for larger ones. RNNs store previous outputs for future computations, making them smarter and more accurate. LSTMs can add more future uses to RNNs by recollecting important information and forgetting irrelevant values. This work explores the use of three cascaded deep learning techniques, multi-layer perception (MLP), RNN, and LSTM, in the classification of voice biomarkers for Parkinson's disease diagnosis[6]. Medical treatments aim to improve life quality for patients with Parkinson's disease (PD). Motor impairments are assessed through neurological examinations or home diaries. The Unified Parkinson's Disease Rating Scale (UPDRS) is a popular rating tool for assessing the progression of PD. Gait movement analysis is an effective choice for detecting PD in a prior state. Wearable gait sensors, such as Ground Reaction Force (GRF) sensors, are popular for PD assessment due to their small size, noninvasive nature, and low cost. Advancements in machine learning can reduce time and workforce problems, and automate gait analysis, making it more efficient for patients and medical staff. This study aims to provide a prognosis solution for PD patients using wearable sensor data. A multistage deep learning approach is proposed, using Convolutional Neural Networks (CNN) and a Locally Weighted Random Forest (LWRF) architecture to predict UPDRS values. The approach aims to predict PD using UPDRS values and outperform previous studies using LWRF models[7].

Research shows that 90% of Parkinson's disease (PD) patients have speech and voice acoustic problems, leading to initial symptom loss. Currently, there is no established treatment, but pharmacological therapies can reduce symptoms. Frequency analysis of voice frequency can track PD progression. Machine learning (ML) approaches are used in the medical sector to diagnose PD in its preclinical phases. The diagnosis involves three phases: pre-processing data, extracting features, and applying classification techniques. The choice of the appropriate classification technique is crucial for PD diagnosis. This review explores the use of ML models trained on sensory data to assist PD patients, their careers, and physicians throughout treatment. It presents key findings from research publications that use IoT technology and sensor installations to improve PD diagnosis and treatment[8]. Health informatics systems are increasingly used for detecting and monitoring diseases, particularly Parkinson's Disease (PD), which is prevalent in people over 60. These systems aim to detect patients with early diagnosis, reduce clinician workload, and recognize the severity of symptoms. Vocal disorders are a common symptom, and speech signal processing techniques are used to extract clinically relevant features. Artificial learning methods like ANN, SVM, Random Forest, and KNN are used for PD classification. The success of these algorithms depends on the quality of features selected from the data[9]. Parkinson's disease (PD) is a neurodegenerative disorder affecting dopaminergic neurons in the substantia nigra region of the brain. Symptoms include bradykinesia, dysarthria, anxiety, depression, sleep behavior disorders, and cognitive impairment. Early diagnosis is crucial for better assessment and quality of life. The Unified Parkinson's Disease Rating Scale (UPDRS) evaluates PD, but its accuracy is limited. Machine and deep learning approaches have been introduced for automated detection and classification, but motor system abnormalities remain the current method of clinical diagnosis due to their subjective nature and lack of established biomarkers[10].

3. Algorithm

3.1 Random Forest(RF):

Random Forest is a powerful machine-learning algorithm used for both classification and regression tasks. It works by creating multiple decision trees from random subsets of the training data, with each tree making independent predictions. These trees are built using a technique called bootstrapping, where random samples of data points are chosen for each tree, which helps reduce overfitting. Additionally, when splitting nodes in each tree, only a random subset of features is considered, further increasing the diversity of the trees. Once the trees are built, Random Forest combines their predictions by taking a majority vote (for classification) or averaging the outputs (for regression). This ensemble approach leads to more accurate and robust models compared to individual decision trees. Random Forest is popular due to its high accuracy, resistance to overfitting, and versatility, making it suitable for a wide range of real-world applications.

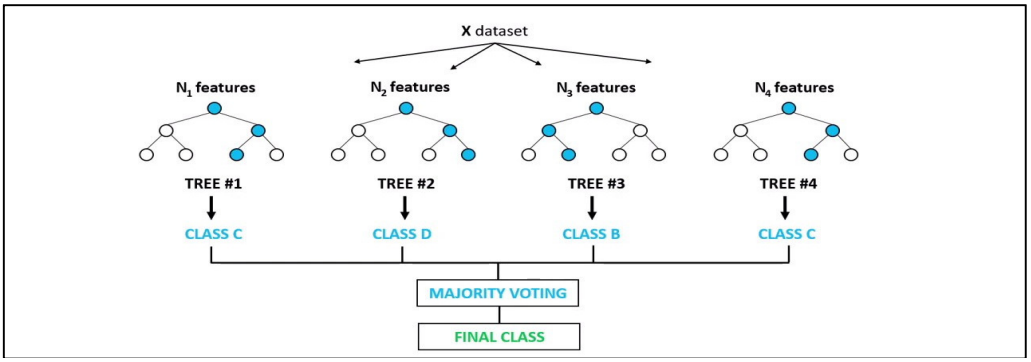

Figure 1: Random Forest(RF)

3.2 Support Vector Regression(SVR):

Support Vector Regression (SVR) offers several advantages, making it a popular choice for regression tasks. It is effective in high-dimensional spaces, meaning it can handle cases where there are many features relative to the number of data points. This makes it particularly useful in fields like bioinformatics or text mining, where feature spaces can be very large. Additionally, SVR is robust to overfitting due to its use of the epsilon margin, which allows the model to focus only on the data points that fall outside of this margin, leading to a simpler, more generalized model. Another key advantage of SVR is its versatility; it can model both linear and non-linear relationships by using different kernel functions, such as polynomial or radial basis function (RBF) kernels. This flexibility allows SVR to handle complex, non-linear data effectively, making it suitable for a wide range of real-world applications

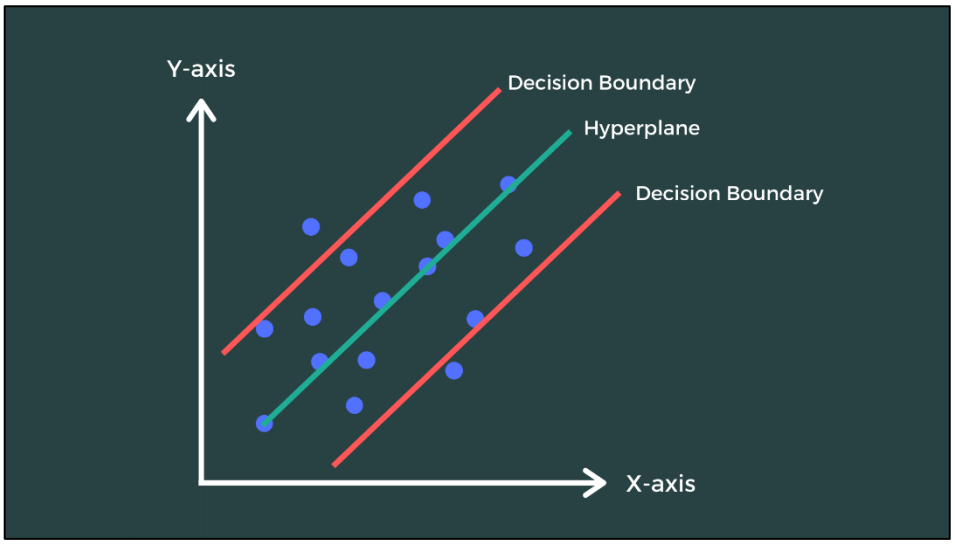

Figure 2: Support Vector Regression(SVR)

4. Methodology:

The above image represents a machine-learning workflow for Parkinson's disease prediction. It begins with a data preprocessing step, where raw data is cleaned and prepared for analysis. The dataset is then split into training and testing subsets. The training data is used to train a machine learning model, which learns patterns associated with Parkinson's disease. The test data is used to validate the model's performance, ensuring accuracy and reliability. Once trained and validated, the model is deployed to predict whether an individual has Parkinson's disease or is normal based on input features. This structured pipeline ensures efficient and accurate disease classification.

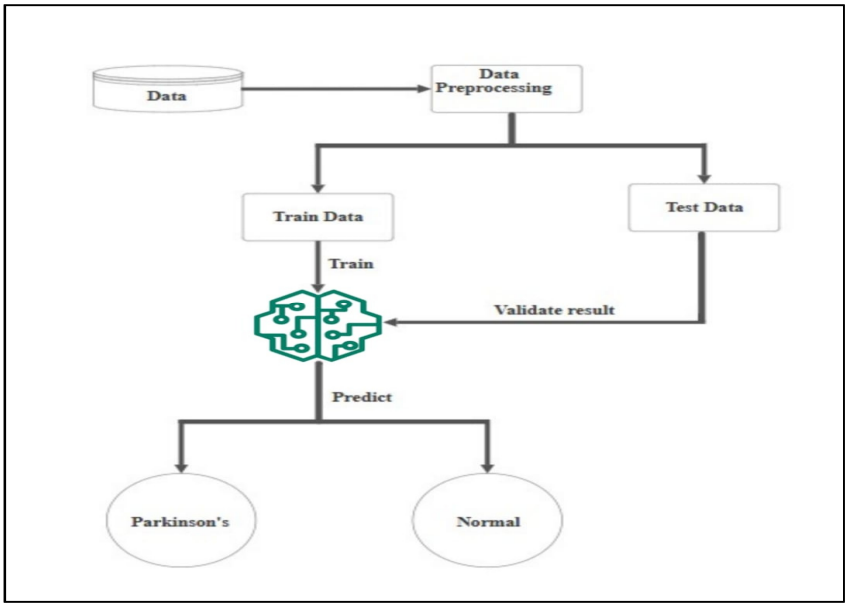

Figure 3: Model Process

The attributes of records are elaborated in Table 1 below:

Field Name	Description	Data Type
subject#	Unique identifier for each subject (1-31)	Numeric
age	Age of Subject	Numeric
sex	Gender of Subject	Categorical
test_time	Time since recruitment into trial	Numeric
motor_UPDRS	Motor UPDRS score	Numeric
total_UPDRS	Total UPDRS score	Numeric
Jitter(%)	Percentage of local variation in fundamental frequency	Numeric
Jitter(Abs)	Absolute jitter	Numeric
Jitter:RAP	Relative average perturbation	Numeric
Jitter:PPQ5	Five-point period perturbation quotient	Numeric
Jitter:DDP	Difference between consecutive differences of fundamental frequency	Numeric
Shimmer	Local variation in amplitude	Numeric
Shimmer(dB)	Shimmer in decibels	Numeric
Shimmer:APQ3	Three-point amplitude perturbation quotient	Numeric
Shimmer:APQ5	Five-point amplitude perturbation quotient	Numeric
Shimmer:APQ11	Eleven-point amplitude perturbation quotient	Numeric

Field Name	Description	Data Type
Shimmer:DDA	Difference between consecutive differences of amplitude	Numeric
NHR	Noise-to-harmonics ratio	Numeric
HNR	Harmonics-to-noise ratio	Numeric
RPDE	Recurrence period density entropy	Numeric
DFA	Detrended fluctuation analysis	Numeric

5. Result :

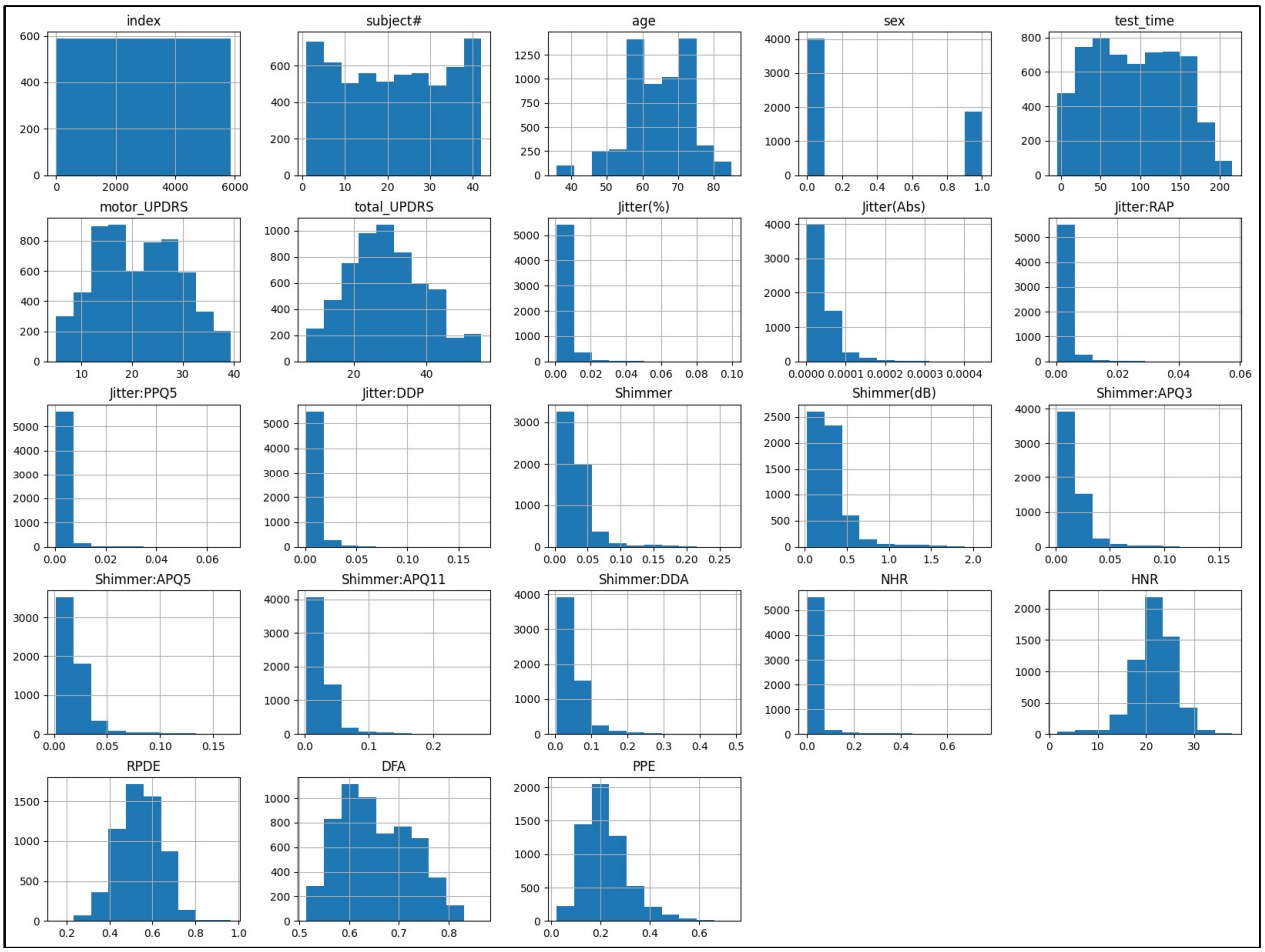

Figure 4: Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) visualizations provide insights into the dataset's structure and distribution. The missing values heatmap confirms no missing data, ensuring data integrity. The age distribution shows a multimodal pattern, indicating varying age groups. The motor_UPDRS and total_UPDRS histograms highlight a skewed distribution, suggesting variability in disease severity among patients. Additionally, feature-wise histograms reveal data distribution patterns and potential skewness, essential for preprocessing and model selection. These analyses help understand data trends, detect outliers, and ensure appropriate feature transformations for accurate predictions.

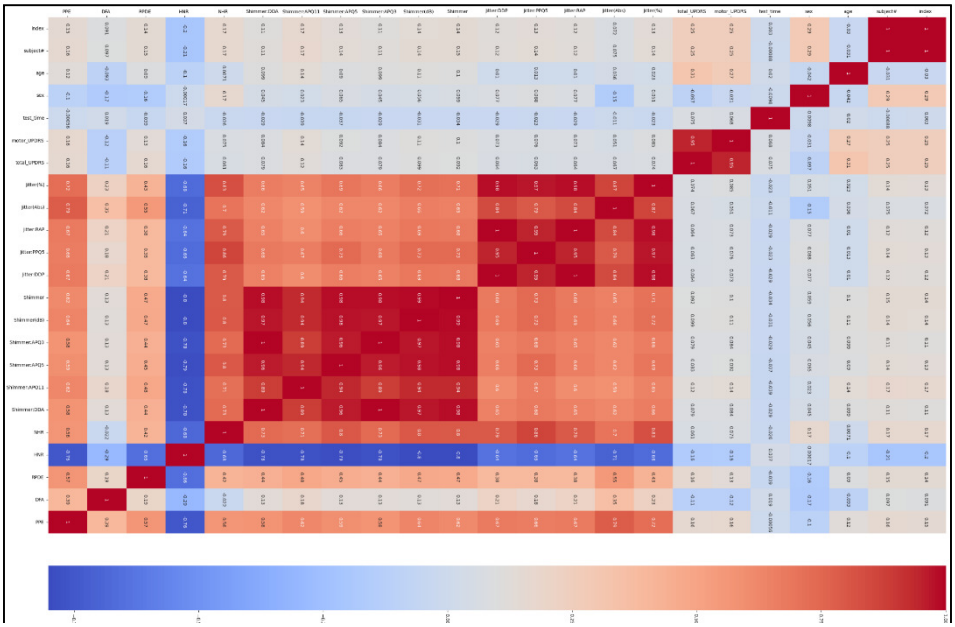

Figure 5: Correlation Matrix

The correlation matrix heatmap visually represents the relationships between numerical features in the dataset, with red indicating strong positive correlations and blue representing negative correlations. Key observations include a strong positive correlation between total_UPDRS and motor_UPDRS, suggesting their interdependence in disease severity. Additionally, jitter and shimmer features exhibit high correlations, while HNR shows a negative correlation with these parameters, aligning with expected speech impairment patterns in Parkinson's disease. This analysis aids in feature selection by identifying redundant variables and improving predictive modeling efficiency.

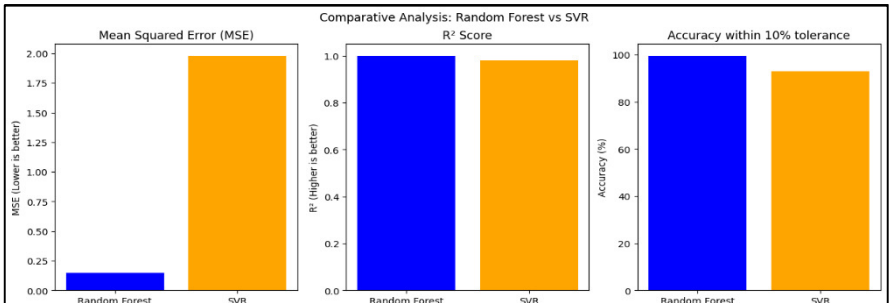

Figure 6: Comparative Analysis(Random Forest vs Support Vector Regression)

The performance comparison between the Random Forest and Support Vector Regression (SVR) models demonstrates that Random Forest outperforms SVR in terms of accuracy and error metrics. The Random Forest model achieved a Mean Squared Error (MSE) of 0.15 and an R^2 score of 1.00, indicating a near-perfect fit. In contrast, the SVR model resulted in a higher MSE of 1.98 and an R^2 score of 0.98, suggesting a slightly lower predictive accuracy. Furthermore, when evaluating accuracy within a 10% tolerance, the Random Forest model exhibited superior performance, achieving 99.57% accuracy compared to SVR's 92.94%. These findings highlight Random Forest as a more robust and precise model for the given dataset.

6. Conclusion:

Parkinson's disease diagnosis is challenging to manage daily. Thus, an effective screening process will be helpful, especially for cases that do not require a visit to a clinic. Symptoms like vocal characteristics, voice recording, speech, and slow movement are valuable and non-invasive diagnostic tools. This paper used machine learning algorithms to diagnose the disease through the patient's voice patient. This is a

practical step to check before meeting with a clinician. A dataset of voices was used as an input to several machine learning models. The results show that the random forest classifier performs with high accuracy. In future work, more datasets of PD patients can be used in order to measure the accuracy of the random forest if the new data is added.

7. Reference:

1. Sriram TV, Rao MV, Narayana GS, Kaladhar DS, Vital TP. Intelligent Parkinson disease prediction using machine learning algorithms. Int. J. Eng. Innov. Technol. 2013 Sep;3(3):1568-72.
2. Krishna PG, StalinDavid D. An Effective Parkinson's Disease Prediction Using Logistic Decision Regression and Machine Learning with Big Data. Turkish Journal of Physiotherapy and Rehabilitation. 2021;32(3):778-86.
3. Wang W, Lee J, Harrou F, Sun Y. Early detection of Parkinson's disease using deep learning and machine learning. IEEE Access. 2020 Aug 12;8:147635-46.
4. Tiwari AK. Machine learning-based approaches for prediction of Parkinson's disease. Mach Learn Appl. 2016 Jun;3(2):33-9.
5. Ahmed I, Aljahdali S, Khan MS, Kaddoura S. Classification of Parkinson's disease based on patient's voice signal using machine learning. Intelligent Automation and Soft Computing. 2022;32(2):705.
6. Chintalapudi N, Battineni G, Hossain MA, Amenta F. Cascaded deep learning frameworks in contribution to the detection of Parkinson's disease. Bioengineering. 2022 Mar 12;9(3):116.
7. Rana A, Dumka A, Singh R, Panda MK, Priyadarshi N, Twala B. Imperative role of a machine learning algorithm for detection +of Parkinson's disease: review, challenges and recommendations. Diagnostics. 2022 Aug 19;12(8):2003.
8. Aşuroğlu T, Oğul H. A deep learning approach for parkinson's disease severity assessment. Health and Technology. 2022 Sep;12(5):943-53.
9. Gunduz H. Deep learning-based Parkinson's disease classification using vocal feature sets. Ieee access. 2019 Aug 20;7:115540-51.
10. Shaban M. Deep learning for Parkinson's disease diagnosis: a short survey. Computers. 2023 Mar 7;12(3):58.

