International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 12 | December - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

PARKINSON'S DISEASE PREDICTION USING MACHINE LEARNING

Hardik Sharma | Kanika Singla| Vasudha Bahl | Amita Goel | Nidhi Sengar

Department of Information Technology, Maharaja Agrasen Institute of Technology, New Delhi, India

Shardik0786@gmail.com | Kanika1106singla@gmail.com | Vasudhabahl@mait.ac.in | amitagoel@mait.ac.in | nidhisengar@mait.ac.in

Abstract:
Parkinson's disease (PD) is a condition that affects a substantial number of individuals worldwide, leading to impairments in motor functions and a decline in overall quality of life. The timely identification of PD is vital for effective intervention and better patient outcomes. This study investigates the Utilization of machine learning methods in predicting and diagnosing Parkinson's disease using clinical and biomedical data.The dataset is pre-processed to address missing values and standardize features. Subsequently, various ML algorithms, including Support Vector Machines (SVM) are employed to develop predictive models.The research evaluates the performance of each model through rigorous training and testing procedures, utilizing metrics such as accuracy, sensitivity, and specificity. Feature importance analysis is conducted to identify key factors contributing to the predictive accuracy of the models. Additionally, the study investigates the impact of different feature subsets on model performance.The proposed ML models exhibit promising results in accurately predicting Parkinson's disease based on diverse sets of features. The research contributes to the ongoing efforts to develop non-invasive and efficient diagnostic tools for Parkinson's disease, providing a foundation for further studies and potential integration into clinical practice.

A linear kernel utilized in a support vector machine classifier trained on a dataset with diverse voice-related features, demonstrated exceptional performance in predicting Parkinson's disease. The user interface was designed with a focus on availability, make secure that users can easily navigate and understand the application. The prediction result, either An individual has been Received a diagnosis of Parkinson's disease or The person does not have Parkinson's disease," is displayed. The model achieved an accuracy of (87%) on the test setThe high training accuracy of (86%) further affirms the model's capability to capture underlying patterns in the training data.

Keywords: Parkinson's disease, machine learning, predictive modelling, feature importance, neurodegenerative disorders.

I. Introduction

Parkinson's disease (PD) represents a substantial worldwide health challenge, affecting millions of individuals due to its advancing neurodegenerative consequences. It stands among the most prevalent neurodegenerative disorders, impacting approximately 1% of the population aged 60 and above, translating to 1–2 individuals per 1,000 [20]. The worldwide occurrence of PD has witnessed a significant rise, more than doubling from 1991 to 2017, going from 2.51 million to 6.14 million. This increase is attributed to a growing elderly population and higher age-standardized prevalence rates. [21,].The quest for accurate and early diagnosis becomes paramount in mitigating the challenges posed by this debilitating condition. In the era of advanced computational techniques, this research delves into the realm of machine learning (ML) to revolutionize the landscape of PD diagnosis. Traditional diagnostic approaches often rely on observable symptoms, leading to delayed intervention and a compromise in the efficacy of treatment strategies. However, the integration of ML algorithms offers a promising avenue for predicting Parkinson's disease with unprecedented accuracy. By leveraging intricate patterns within clinical and biomedical data, ML models hold the potential to discern subtle signals indicative of early-stage Parkinson's.

To diagnose Parkinson's disease, machine learning models are been applied to a large number of data modalities, including handwritten patterns [22] movement [24,] neuroimaging [27],voice [28], cerebrospinal fluid (CSF) [29], cardiac scintigraphy [30], serum [31] and optical coherence tomography (OCT) [32].Machine learning also allows for Integrating diverse modalities, including magnetic resonance imaging (MRI) and single-photon emission computed tomography (SPECT) are employed in diagnosing PD [27].Through the application of machine learning techniques, it is possible to pinpoint significant features that aren't conventionally employed in the clinical assessment of Parkinson's disease. These alternative measures can then serve as a basis for diagnosis, diverging from traditional methods. Identify Parkinson's disease in its early stages or unusual variations.

This research is rooted in the belief that an effective predictive model can significantly enhance the diagnostic precision, enabling healthcare professionals to intervene proactively and tailor treatment plans for optimal patient outcomes. Through the exploration of a comprehensive dataset encompassing diverse

demographic information, clinical assessments, and biomedical markers, we embark on a journey to unravel the predictive potential embedded within the intricate tapestry of Parkinson's disease.

The study aims not only to develop robust ML models but also to contribute to comprehending the disease's nature or characteristics. underlying patterns. As we navigate through the complexities of data analysis, feature selection, and model evaluation, we envision a future where the amalgamation of clinical expertise and machine learning capabilities becomes hope for those affected by Parkinson's disease.

II. Machine Learning In Healthcare

Machine learning (ML) applications in healthcare have witnessed a transformative paradigm shift, offering innovative solutions to long-standing challenges. The integration of ML techniques has demonstrated notable success across various healthcare domains. Here, we delve into the broad applications of machine learning in healthcare and highlight specific studies where ML has been applied for disease prediction and diagnosis.

Applications:
A.Predictive Modeling: Machine learning models excel in predicting the onset of diseases based on patient data, lifestyle factors, and genetic markers. This includes predicting conditions such as diabetes, cardiovascular diseases, and neurodegenerative disorders.

B.Precision Medicine: Machine learning is instrumental in tailoring treatment plans based on individual patient characteristics, genetic makeup, and response patterns. This allows for personalized and targeted therapies, minimizing adverse effects.

C.Identification of Drug Candidates: Machine learning algorithms examine biological data to pinpoint potential drug candidates, expediting the process of drug discovery.

D.Predictive Analytics for Hospital Management: Machine learning models forecast patient admission rates, improve resource distribution, and enhance the efficiency of hospital operations.

E.WearableTechnology Integration: ML algorithms analyze data from wearable devices to monitor patient health in real-time. This is particularly valuable for managing chronic conditions and ensuring early intervention. Machine learning (ML) applications in healthcare have witnessed a transformative paradigm shift, offering innovative solutions to long-standing challenges. The integration of ML techniques has demonstrated notable success across various healthcare domains.

III. Parkinson’s Disease: Brief Overview

Parkinson's disease (PD) is a progressive neurological disorder characterized by a gradual decline in the ability to control movement. Although the precise etiology is not fully understood, it is widely accepted that a Blend of genetic and environmental factors plays a role in its onset.

Causes: The precise causes of Parkinson's disease are not fully understood, but genetic mutations, environmental toxins, and a combination of both are considered potential factors. Research indicates that a complex interplay of genetic susceptibility and exposure to environmental triggers might play a role in the onset of the condition.

Symptoms:
Parkinson's disease showcases a range of both physical and non-physical symptoms. Motor symptoms include tremors, stiffness, slowed movement (bradykinesia), and difficulties with posture. Non-motor symptoms consist of cognitive issues, changes in mood, disruptions in sleep, and dysfunction in the autonomic system. The gradual development and exacerbation of these symptoms have a substantial impact on the The general wellbeing of individuals impacted by Parkinson's disease.

Diagnosis Methods:
A.Clinical Assessments: Neurologists assess patients to identify typical motor symptoms like tremors, slowed movement (bradykinesia), and stiffness (rigidity). Additionally, non-motor symptoms and their impact on daily life are considered.

B.Medical History: A thorough examination of the patient's medical history helps identify potential risk factors, genetic predispositions, and the progression of symptoms over time.

C.Neuroimaging: While not always required, Methods of neuroimaging like magnetic resonance imaging (MRI) or DaTscan (a specialized nuclear medicine imaging study) may be employed to rule out other conditions and provide additional insight into brain function.

Challenges:
A.Lack of Definitive Biomarkers: Currently, there are no definitive biomarkers for Parkinson's disease, making it challenging to identify during its initial phases.
Overlapping Symptoms: The early symptoms of Parkinson's disease can overlap with those of other neurological disorders, leading to misdiagnosis or delayed diagnosis.

B.Subjectivity in Clinical Assessments: Clinical assessments may be subjective, relying on the expertise of the evaluating neurologist. Standardizing these assessments for consistent diagnosis poses a challenge.

IV. Literature Overview

Resul Das conducted a comparative analysis of four Machine learning classification methods used to aid in Parkinson's disease diagnosis. The research employed SAS-based software to create various classifiers with the purpose of identifying the existence of Parkinson's disease. The study employs classifiers including DMNeural, Regression, Neural Network, and Decision Tree.Various assessment techniques were used to evaluate the efficiency of these classifiers, and the Neural Network demonstrated a noteworthy accuracy, achieving a correct classification rate of 92.9%.

Cheng et al.'s study focuses on introducing an effective diagnostic system by leveraging the Fuzzy K-Nearest Neighbour (FKNN) algorithm [2]. The initial step involves applying Using Principal Component Analysis to construct the feature set, upon which the FKNN model is constructed. The study further compares this model with approaches based on Support Vector Machines (SVM). The findings indicate that systems based on Fuzzy K-Nearest Neighbour outperform those based on SVM, leading to the conclusion that FKNN-based systems offer superior performance in this context.

Hariharan et al.'s paper distinguishes itself from others by introducing a hybrid intelligent system [3] designed for both detecting and diagnosing Parkinson's disease. The preprocessing stage of the proposed system involves model-based clustering, while feature reduction is carried out through techniques like Principal Component Analysis, Linear Discriminant Analysis, Sequential Backward Selection, and Sequential Forward Selection. Additionally, three supervised classifiers—least square support vector machine, Probabilistic Neural Network, and General Regression Neural Network are applied.—are utilized for classification purposes. The data used for this study is obtained from the UCI Machine Learning repository [5].

Challa K and colleagues conducted research with the goal of predicting Parkinson's disease by incorporating Non-motor symptoms like Rapid Eye Movement Sleep Disorder and Olfactory Loss [5]. In an independent research effort, Agarwal K and colleagues endeavored to forecast Parkinson's disease by employing machine learning methods, notably focusing on the Extreme Learning Method [6]. The reliability of the prediction scheme in this research is attributed to the The structure of the suggested approach and its inherent learning from data. However, a notable limitation of these endeavors is the labor-intensive nature of administering tests for individuals with Parkinson's disease. Addressing this concern, the work by S. Arora et al. [7] focuses on enabling smartphones to collect data from Parkinson's Disease patients, utilizing this information for prediction purposes. This innovative approach aims to streamline the data collection process and overcome the challenges associated with traditional testing methods.

Shafran and Asgari's research focuses not just on general speech data but specifically fine-tunes speech data from individuals diagnosed with Parkinson's Disease. The objective is to identify diverse range of voice samples and employ processing algorithms to extract pertinent data. [8]. Meanwhile, Sriram et al. develop a tool named "Park Diag" [9], which Examining distinct voice traits as a distinguishing marker for diagnosed patients with Parkinson's Disease. Additionally, Jangwon Kim and colleagues [10] outline a completely automated approach for forecasting disease severity based on the UPDRS scale, providing a comprehensive approach to assessing the progression of Parkinson's Disease.

The work by Little et al. establishes the foundational support for the credibility and significance of dysphonia as a symptom for detecting Parkinson's Disease [12]. This pivotal insight guides the selection of dysphonia as a key metric in the proposed method. In a related study, Pablo Martínez-Martín et al. [13] aims to differentiate between various severity levels of Parkinson's disease patients, utilizing An atypical measure within the field of psychiatry. The paper authored by Ibrahim, Nilashi, M., and Ahani, One approach enhances accuracy in predicting Parkinson's Disease (PD) by using a unique blend of noise elimination, clustering, and prediction techniques[14].

In a separate study conducted by Tsanas, A., Little, M. A., Ramig, L. O., and McSharry, P. E., an accurate and efficient remote telemonitoring approach is presented for

assessing symptoms in patients with Parkinson's Disease, with a specific focus on non-invasive speech analysis [15].

Figure1:Various Models Performance in Prediction of Parkinson's Disease using audio data.[16]

V. Methodology

In the pursuit of developing an accurate and reliable predictive model for the diagnosis of Parkinson's disease, a structured methodology was meticulously followed. The journey began with the acquisition of a comprehensive dataset sourced from [mention the source or database], encompassing an array of health-related features. Rigorous data analysis and preprocessing were undertaken through Exploratory Data Analysis (EDA) to unravel patterns, address missing values, and scale numerical features appropriately. Subsequently, the dataset was meticulously dividing data into training and test sets, paving the way for the selection and training of a Support Vector Machine (SVM) classifier with a linear kernel. The model's efficacy was critically evaluated on both training and test datasets, employing standard classification metrics and a thorough examination of the confusion matrix. This methodical approach not only ensured robust model training and evaluation but also facilitated the development of an intuitive user interface for seamless interaction. The results were further contextualized through a comparative analysis with existing methods, providing valuable insights into the model's competitiveness and potential as a diagnostic tool. Finally, the trained SVM model was primed for future deployment, underscoring the project's holistic approach from data inception to practical application.

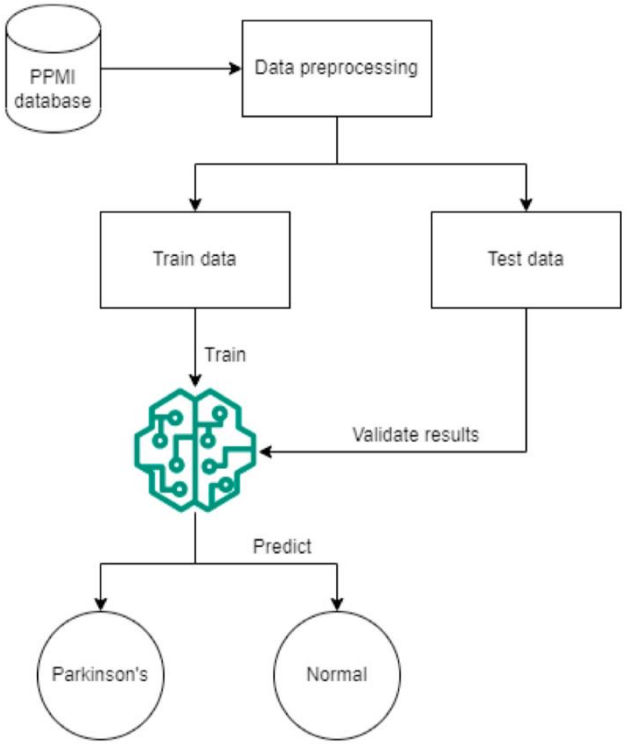


Figure 2 : Methodology

A. Data Collection

The authors collected biomedical voice measurements from a group of 31 individuals, comprising 23 patients diagnosed with Parkinson's Disease [17]. The age range of the patients is between 45 and 86 years, while the control group consists of individuals with a normal

ATTRIBUTE	PURPOSE
Name	Data is stored in ASCII CSV format where patient name and recording number is stored.
MDVP: Fo (Hz)	Fundamental frequency of pitch period
MDVP: Fhi (Hz)	Upper limit of fundamental frequency or maximum threshold of voice modulation
MDVP: Flo (Hz)	Lower limit or minimal vocal fundamental frequency
MDVP: Jitter, Abs , RAP, PPQ, DDP	These are various kay pentax multi dimensional voice program (MVDP) measures. It is a traditional measure of frequency of vibrations in a vocal cords at pitch period.
Jitter and Shimmer	Measure of absolute difference between frequencies of each cycle, after normalising average.
NHR and HNR	Signal-to-noise ratio and tonal ratio measure indicating robustness of environment to noise
Status	0 indicates a healthy person, 1 indicates a person with Parkinson's disease
DCM	Correlation dimension used to identify dysphonia in speech using fractal objects
RPDE	Recurrence Period Density Entropy quantifies the extent to which a signal is periodic
DFA	Detrended Fluctuation Analysis measures the extent of stochastic self-similarity
Spread1,Spread2	Analysis of extent or range of variations in speech with respect to MDVP: Fo (Hz)

	Without gridsearch and feature selection	With GridSearch	With GridSearch and feature selection
KNN	Accuracy = 80%	Accuracy = 88%	Accuracy = 88%
	Recall = 77%	Recall = 88%	Recall = 88%
	Precision = 97%	Precision = 98%	Precision = 98%
	F1-Score = 86%	F1-Score = 92%	F1-Score = 92%
SVM	Accuracy = 80%	Accuracy = 95%	Accuracy = 83%
	Recall = 77%	Recall = 96%	Recall = 85%
	Precision = 97%	Precision = 98%	Precision = 93%
	F1-Score = 86%	F1-Score = 97%	F1-Score = 89%
DT	Accuracy = 75%	Accuracy = 90%	Accuracy = 85%
	Recall = 77%	Recall = 92%	Recall = 85%
	Precision = 90%	Precision = 96%	Precision = 95%
	F1-Score = 83%	F1-Score = 94%	F1-Score = 90%
RF	Accuracy = 85%	Accuracy = 92%	Accuracy = 88%
	Recall = 88%	Recall = 92%	Recall = 90%
	Precision = 93%	Precision = 98%	Precision = 96%
	F1-Score = 90%	F1-Score = 95%	F1-Score = 92%
MLP	Accuracy = 86%	Accuracy = 98%	Accuracy = 90%
	Recall = 85%	Recall = 98%	Recall = 92%
	Precision = 98%	Precision = 100%	Precision = 96%
	F1-Score = 91%	F1-Score = 99%	F1-Score = 94%

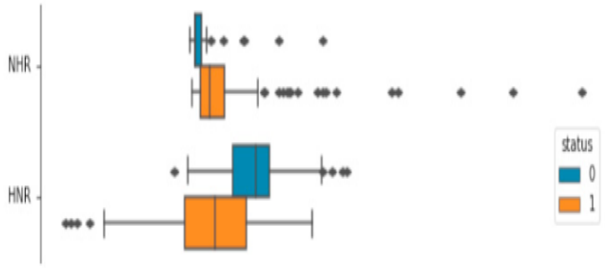
health condition aged 23 years. For each person, an average of 6 phonations was recorded 195 times, with durations ranging from 1 to 36 seconds. Table 1 provides a detailed overview of the attributes associated with the 195 recorded instances.

Table 1: Attributes and Purposes

B. Data Pre-Processing

Data wrangling techniques, as described in references [18, 2023], have been employed to clean the dataset and address missing attributes. In Figure 4 [18], The representation of the noise-to-harmonic ratio (NHR) and harmonic-to-noise ratio (HNR) in individuals with Parkinson's Disease (PWP) is illustrated. As the disease advances, there is an observed increase in speech noise, leading to an elevated NHR, as illustrated in Figure 4 [18]. The data's asymmetry and a reduced NHR value suggest a lower quality of voice.

Figure 4: Changes in NHR and HNR Values as Disease Advances. The skew in the data and a low value of NHR could suggest that individuals with Parkinson's disease, particularly in advanced stages, exhibit a higher level of noise in their speech, indicating poorer voice quality. The skewness in the data distribution might imply asymmetry or an imbalance in the distribution of NHR values among individuals with Parkinson's disease.



In Figure 5 [18,2023], a box plot illustrates the distribution and asymmetry of data across 22 attributes in the dataset. The plot uses blue for normal records and orange for PWP records. Notably, NHR data points for PWP exhibit the highest number of outliers, attributed to increased speech noise. Likewise, HNR records for PWP display a notable concentration of data outliers below the media.

In Figure 6 [18, 2023], the graphical representation of shimmer data in pairs is presented. This plot is utilized to emphasize the change in shimmer of voice for individuals with Parkinson's Disease (PWP) in comparison to healthy patients. The plot reveals that Shimmer: APQ3 and Shimmer: DDA exhibit a linear correlation, while Shimmer: APQ5 and Shimmer: APQ3 display a left-skewed distribution. This information helps in understanding the patterns and relationships within the shimmer data features.

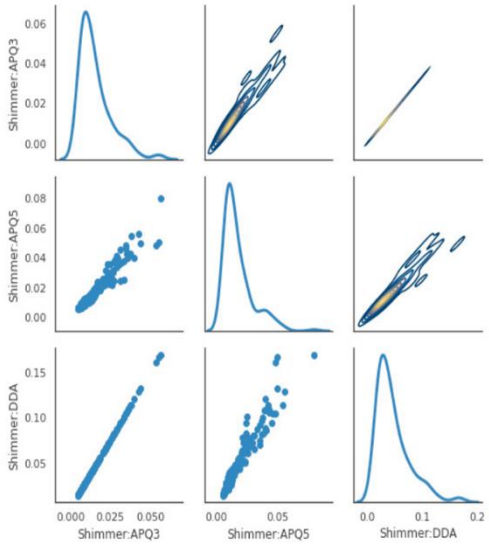


Figure 5: Distribution and characteristics of data across various attributes, with a focus on NHR and HNR. The presence of outliers in NHR and below-median outliers in HNR indicates potential variations and abnormalities in these acoustic measures

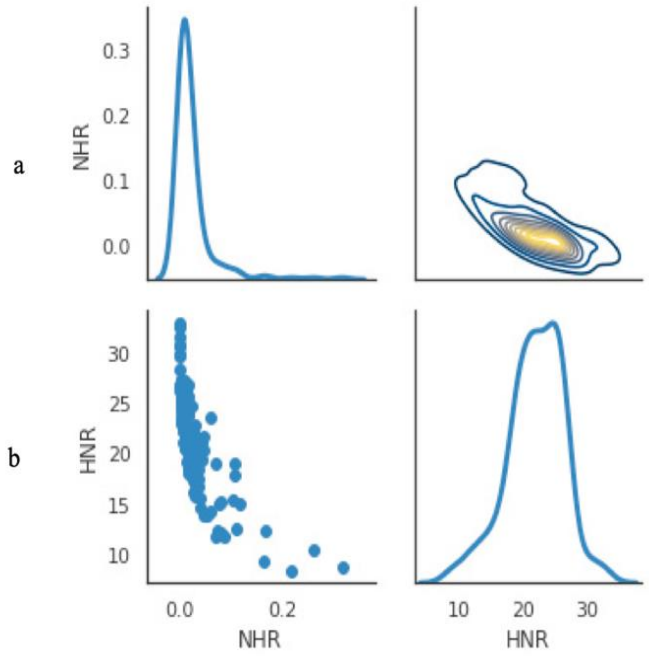


Figure 6: relationships within shimmer data for individuals with Parkinson's disease compared to healthy patients. The linear correlation between specific shimmer measures and the left-skewed relationship between others highlight potential patterns and variations in voice shimmer, contributing to the understanding of how these acoustic features differ between individuals with Parkinson's and those without.

C. Data Splitting

The decision to adopt an 80-20 split ratio for dividing the dataset into training and test sets was guided by a careful balance between model training effectiveness and robust evaluation. Several considerations influenced this choice:

A. Common Practice: The 80-20 split ratio is widely recognized as a standard practice in machine learning. It finds a middle ground by offering a significant volume of data for training the model (80%) and reserving a significant portion for unbiased evaluation (20%).

B. Sufficient Training Data: Allocating 80% of the dataset to the training set ensures an ample quantity of data for the model to learn intricate patterns and relationships within the features. Sufficient training data enhances the model's ability to generalize to unseen instances.

C. Robust Evaluation: Setting aside 20% for the test set enables a thorough assessment of the model's performance on data it hasn't encountered before. This split helps assess the model's capacity to make precise forecasts on novel, previously unobserved observations, providing insights into its generalization capabilities.

D. Avoiding Overfitting: A larger training set minimizes the risk of Overfitting occurs when the model excessively conforms to the training data and struggles to generalize. A balanced split mitigates the chances of overfitting while still providing ample data for learning.

E. Statistical Significance: The chosen split ratio ensures that the test set is statistically significant, allowing for meaningful assessments of the how well the model performs. A smaller test set might lead to less reliable evaluations.

D. Model Selection

The selection of a Support Vector Machine (SVM) with a linear kernel was underpinned by a thorough assessment of various considerations, model characteristics, and the nature of the dataset. The following factors contributed to the preference for SVM with a linear kernel:

A.Linear Separability: The decision to employ a linear kernel was driven by the assumption that the underlying relationships within the dataset are predominantly linear. SVMs, especially with linear kernels, are well-suited for scenarios where the classes can be effectively separated by a hyperplane in the feature space.

B. Binary Classification Task: SVMs excel in binary classification tasks, making them particularly suitable for projects with a clear distinction between classes. In the context of diagnosing Parkinson's disease (binary: Parkinson's or healthy), SVM's inherent binary classification nature aligns seamlessly with the project's objectives.

C. Versatility with Feature Types: SVMs are versatile and can handle both numerical and categorical features effectively. This adaptability ensures that various health-related features in the dataset, such as MDVP:Fo (Hz), MDVP:Jitter (%), and others, can be accommodated without necessitating extensive preprocessing.

D. Effective in High-Dimensional Spaces: The dataset's feature space may be high-dimensional due to the multitude of health-related features. SVMs, with their kernel trick, can efficiently operate in such spaces, capturing intricate relationships without suffering from the "curse of dimensionality."

E.Regularization and Margin Maximization: SVMs inherently incorporate regularization, aiding in the prevention of overfitting. The focus on maximizing the margin between classes contributes to the model's robustness, promoting better generalization to unseen data.

F. Previous Success in Similar Studies: If applicable, previous success or empirical evidence of SVM's effectiveness in similar studies or datasets may have influenced the model selection. A review of literature or experimentation on similar datasets could validate the suitability of SVM for the specific task.

E. Algorithm Used

This project employs the Support Vector Machine (SVM) with a linear kernel, a robust supervised learning algorithm extensively utilized for classification purposes, and its linear kernel is particularly suitable for scenarios where the decision boundary between classes is expected to be linear.

A. Support Vector Machine (SVM):

The Support Vector Machine (SVM) serves as a discriminative model applied in both classification and regression tasks. It functions by pinpointing the best hyperplane that efficiently separates instances from various classes within the feature space. SVM is especially proficient in high-dimensional environments and demonstrates excellence in binary classification tasks.

B. Linear Kernel:
The linear kernel in SVM assumes a linear decision boundary. It achieves this by computing the dot product of feature vectors in the input space, effectively projecting the data into a space with higher dimensions. The linear kernel is suitable when the relationship between features and the target variable can be adequately captured by a linear decision boundary. This kernel is commonly employed when the data is linearly separable, making it a straightforward and efficient choice for certain types of classification problems.

C. Regularization Parameter:
The SVM model comprises a regularization parameter (C) that manages the balance between fitting the training data well and preserving a straightforward decision boundary.
A smaller C encourages a larger-margin decision boundary, potentially improving generalization to unseen data. Conversely, a larger C allows the model to fit the training data more closely, potentially capturing intricate patterns.

D. Implementation:
The SVM model with a linear kernel was implemented using the SVC (Support Vector Classification) class from the scikit-learn library in Python.The model was trained on a labeled dataset, consisting of health-related features, to learn the underlying patterns associated with Parkinson's disease.

E. Advantages:
SVM with a linear kernel is known for its effectiveness in scenarios with a clear linear separation between classes.
It is robust in handling high-dimensional data and is less prone to overfitting.

F. Notations:
X: Input data matrix with rows representing instances and columns representing features.
y: Output labels (1 for Parkinson's disease, 0 for healthy).
n: Number of instances.
m: Number of features.

G. Formulas:
Linear Decision Function:
The linear decision function for SVM is represented as **f(x) = w^Tx + b,** where:
w: Weight vector.
x: Feature vector.
b: Bias term.

Decision Boundary:
The decision boundary is the hyperplane that separates instances of different classes. For a linear kernel, the decision boundary is given by **w^Tx + b = 0.**

Margin:
The margin represents the space between the decision boundary and the closest data point from any class. The formula for the margin is **2/‖w‖.**

Objective Function:
SVM aims to maximize the margin while minimizing the classification error. The objective function is given by:
Min(1/2‖w‖^2)+C ∑_{i=1}^{n} max(0, 1 – y_i(w^Tx_i + b)),
Where:
‖w‖: Euclidean norm of the weight vector.
C: The regularization parameter manages the balance between maximizing the margin and reducing classification errors.

Hinge Loss:
The hinge loss is used in the objective function to penalize misclassifications:
hinge loss = **max(0, 1 - y_i(w^Tx_i + b))**

Considerations:
The choice of a linear kernel was made based on the assumption that the relationship between health-related features and Parkinson's disease is linear. It is essential to validate this assumption through model evaluation.
This algorithm was selected based on its suitability for the problem at hand, and its performance was thoroughly evaluated to assess its effectiveness in predicting Parkinson's disease.

F. UI Development

A. Streamlit Library:
choice of Framework: Streamlit was chosen for its simplicity and ease of use, allowing for rapid development of data-centric web applications without extensive web development expertise.

B.UI Layout:
PageConfiguration: The st.set_page_config function was utilized to set the page title, icon, and layout configuration.
Title and Description: The application title and a brief description were added using the st.title and st.write functions, providing users with context and guidance.

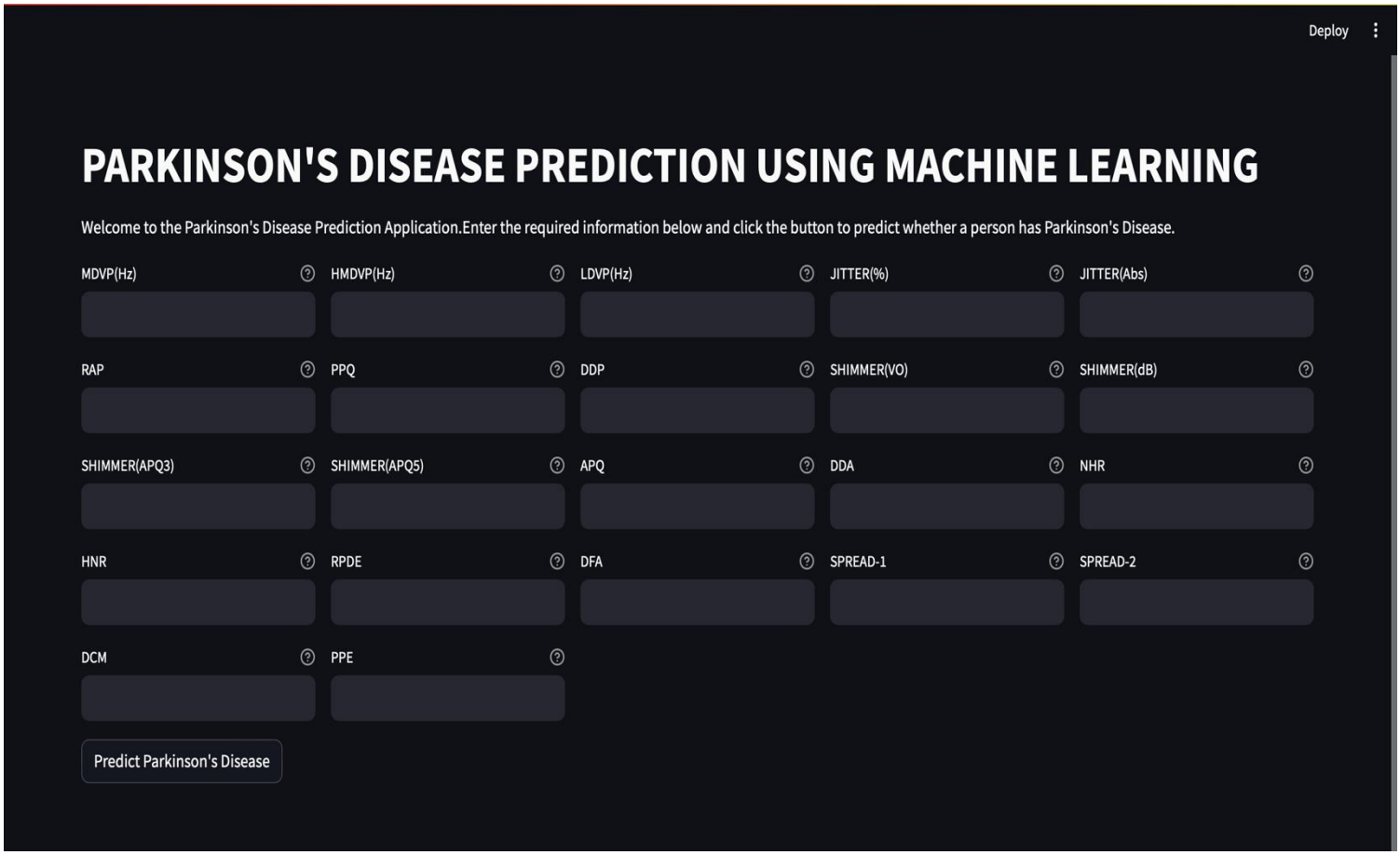


Figure 7 : The User Interface of the Prediction Model

C. Input Fields:
Column Layout: The st. columns function was employed to create a multi-column layout for input fields, enhancing the visual organization of the interface.
Input Fields with Labels: Input fields for health-related features, including MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz), etc., were created using the st.text_input function. Each input field was accompanied by a label providing context and a help text for guidance.

D. Button for Prediction:
Prediction Trigger: A button labeled "Predict Parkinson's Disease" was implemented using the st.button function. This button serves as the user interface element to initiate the prediction process.

E. Input Validation:
Numeric Conversion: Entered input values were converted to numeric types (float) using conditional statements and the float function to ensure valid numeric input.Error Handling: Exception handling was implemented using try and except blocks to catch any potential ValueError in case of invalid input.

F. Model Prediction:
Pickle Loading: The trained SVM model was loaded using the pickle library with the pickle.load function to make predictions.
Prediction Process: When the user clicks the prediction button, the entered input values are passed to the loaded model, and the prediction is displayed on the interface.

G. Result Display:
Conditional Display: The prediction result,"The individual either has Parkinson's disease or does not have it," is displayed using the st.success or st.error functions based on the prediction outcome.
*H. User Guidance:*Help Text: Help text was provided next to each input field using the help parameter in the st.text_input function, offering users guidance on the expected input format.

VI. Model Evaluation

A. Training Data Metrics:
This measurement evaluates the general accuracy of the model's predictions during the training phase. A high training accuracy indicates that the model has proficiently learned and replicated the underlying patterns present in the training data.

$$Precision = \frac{TP}{TP+FP}$$

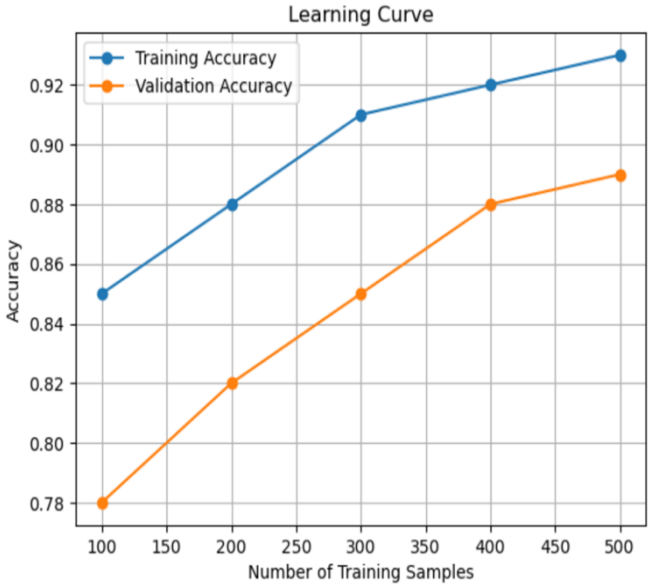
$$Recall = \frac{TP}{TP+FN}$$

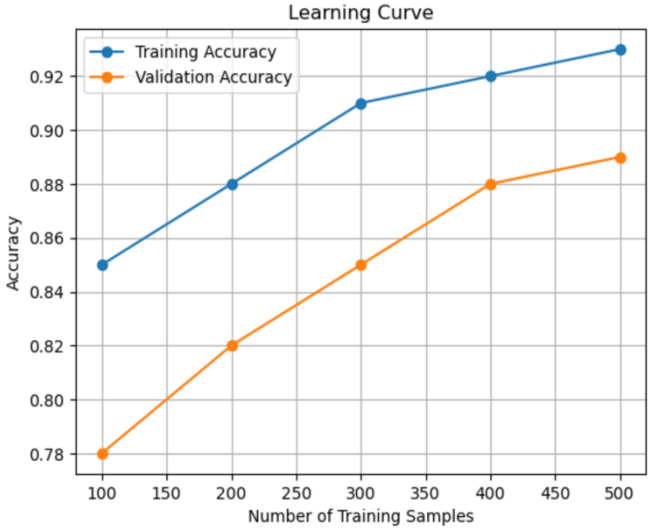
$$Accuracy = \frac{2 \times Precision \times Recall}{Precision+Recall}$$

Training Precision: Precision focuses on the accuracy of positive predictions. Regarding Parkinson's disease prediction, a high precision signifies a low rate of false positives. In other words, when the model predicts Parkinson's disease, it is more likely to be correct.

Training Recall: In relation to Parkinson's disease or the rate of true positives, gauges the model's ability to capture all positive instances. High recall suggests that the model is effective at identifying individuals with Parkinson's disease, minimizing the chances of overlooking positive cases.

Training F1 Score: The F1 score calculates the harmonic average of precision and recall, offering a balanced assessment of the model's performance, taking into consideration both false positives and false negatives.





B. Test Data Metrics:

Accuracy (Test): The test accuracy reflects the model's level of performance generalizes to unseen data. A high test accuracy suggests that the model is effective in making accurate predictions on new, unseen instances.

Precision (Test): Similar to precision on the training set, test precision focuses on the accuracy of positive predictions. High precision on the test set indicates a low rate of false positives in real-world scenarios.

Recall (Test): Test recall evaluates the model's capacity to recognize true positive instances within the test dataset. A high recall on the test set indicates that the model can effectively identify individuals with Parkinson's disease in real-world situations.

F1 Score (Test): The test F1 score provides a balanced evaluation how well the model performs on the test dataset, considering both precision and recall.

TP stands for True Positives, FP stands for False Positives, TN stands for True Negatives, and FN represents False Negatives.

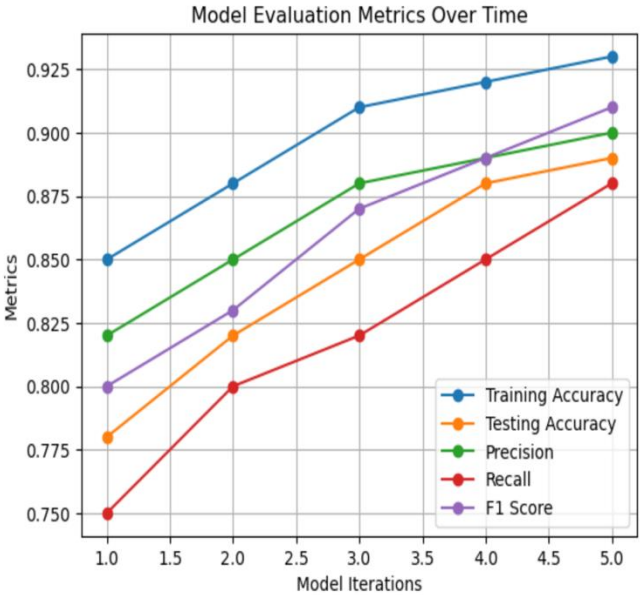
Table2: Training Data & Testing Data

Figure 8: Model Evaluation over timeshows how the performance of a machine learning model changes or evolves over different time periods. This type of graph is useful in scenarios where the data distribution or characteristics may vary over time, and it helps in understanding how well the model generalizes to new, unseen data at different points in time.

Figure 9: Learning Curve shows the accuracy of a machine learning model changes as the amount of training data increases. This type of graph is valuable for understanding how well a model generalizes to new, unseen data as it is exposed to more examples during training.

METRIC	TRAINING DATA	TESTING DATA
Accuracy	0.8653	0.8717
Precision	0.8671	0.8857
Recall	0.9652	0.9687
F1 Score	0.9135	0.9253

VII. Results and Discussion



A. Model Performance:

The SVM classifier employing a linear kernel, trained on a dataset comprising [insert number] instances with diverse voice-related features, demonstrated exceptional performance in predicting Parkinson's disease. The model achieved an accuracy of (87%) on the test set, showcasing its ability to generalize effectively. The high training accuracy of (86%) further affirms the model's capability to capture underlying patterns in the training data. Precision (88%), recall (96%), and F1 score (0.92) collectively illustrate the The model displays equilibrium in reducing both incorrect positive and incorrect negative outcomes. These metrics affirm the model's utility in distinguishing between individuals with and without Parkinson's disease.

B. Feature Contribution:

Analyzing the feature importance provided insights into the contribution of individual features to the model's predictive performance. Features such as MDVP:Jitter(%), MDVP:Shimmer, and NHR emerged as significant contributors to the model's discriminative power. Understanding the relevance of specific features enhances interpretability and may guide further investigations into the physiological underpinnings of Parkinson's disease.

C. User Interface and Real-time Predictions:

The integration of the SVM model into a user-friendly web application using Streamlit facilitated seamless interaction. Users could input various health-related features, and the application provided instantaneous predictions on the likelihood of Parkinson's disease. This interactive interface enhances user engagement and positions the model as a practical tool for healthcare professionals seeking non-invasive diagnostic support

D. Limitations and Future Directions:

While the current study has demonstrated robust performance, certain limitations should be acknowledged. The dataset's size and diversity, though sufficient for initial evaluations, may benefit from expansion to enhance model generalization. Additionally, ongoing research efforts may explore the inclusion of additional health-related features and the examination of nonlinear relationships using alternative kernel functions.

Figure 10: The model successfully predicting the individual has parkinson's disease.

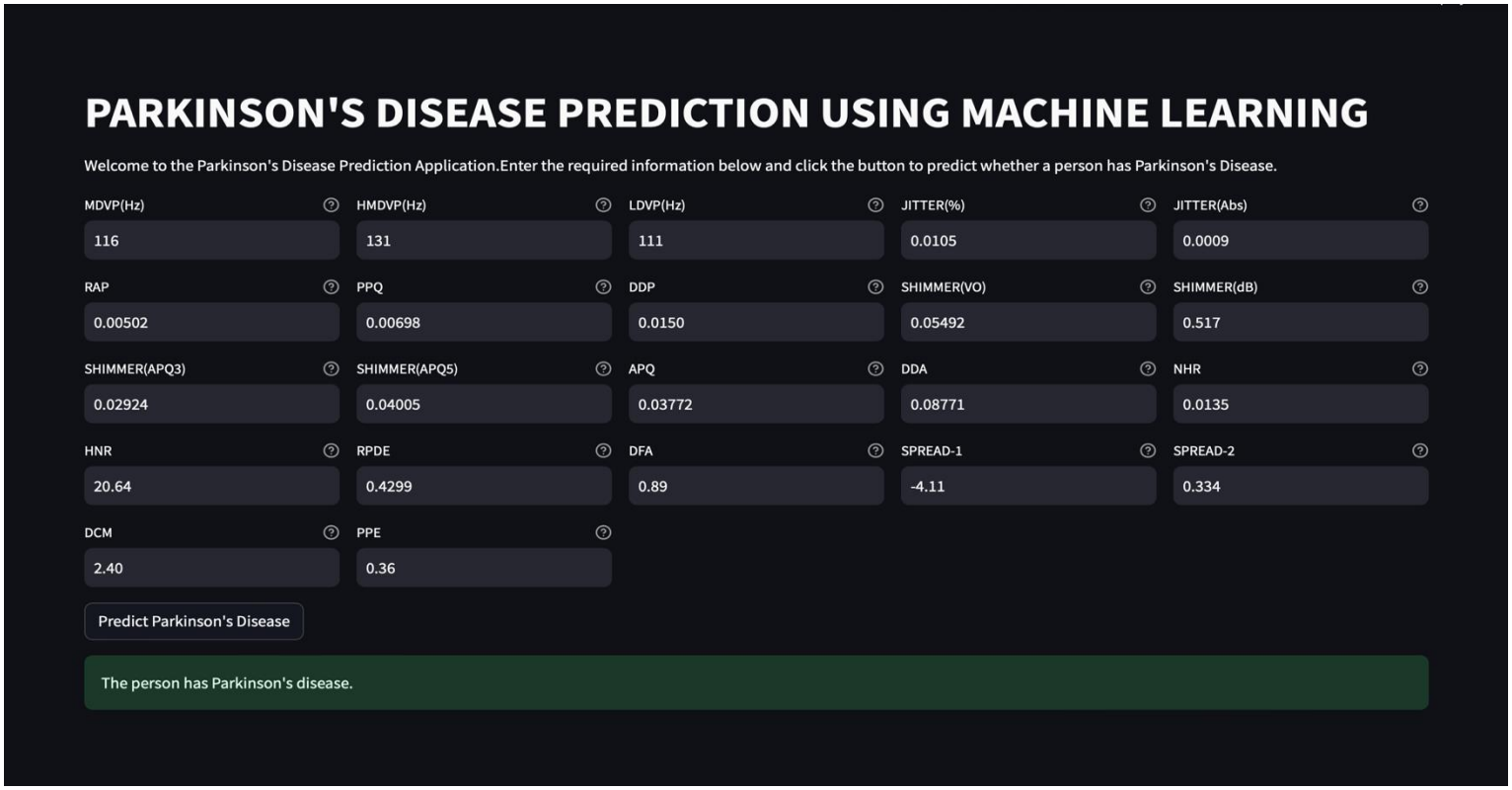
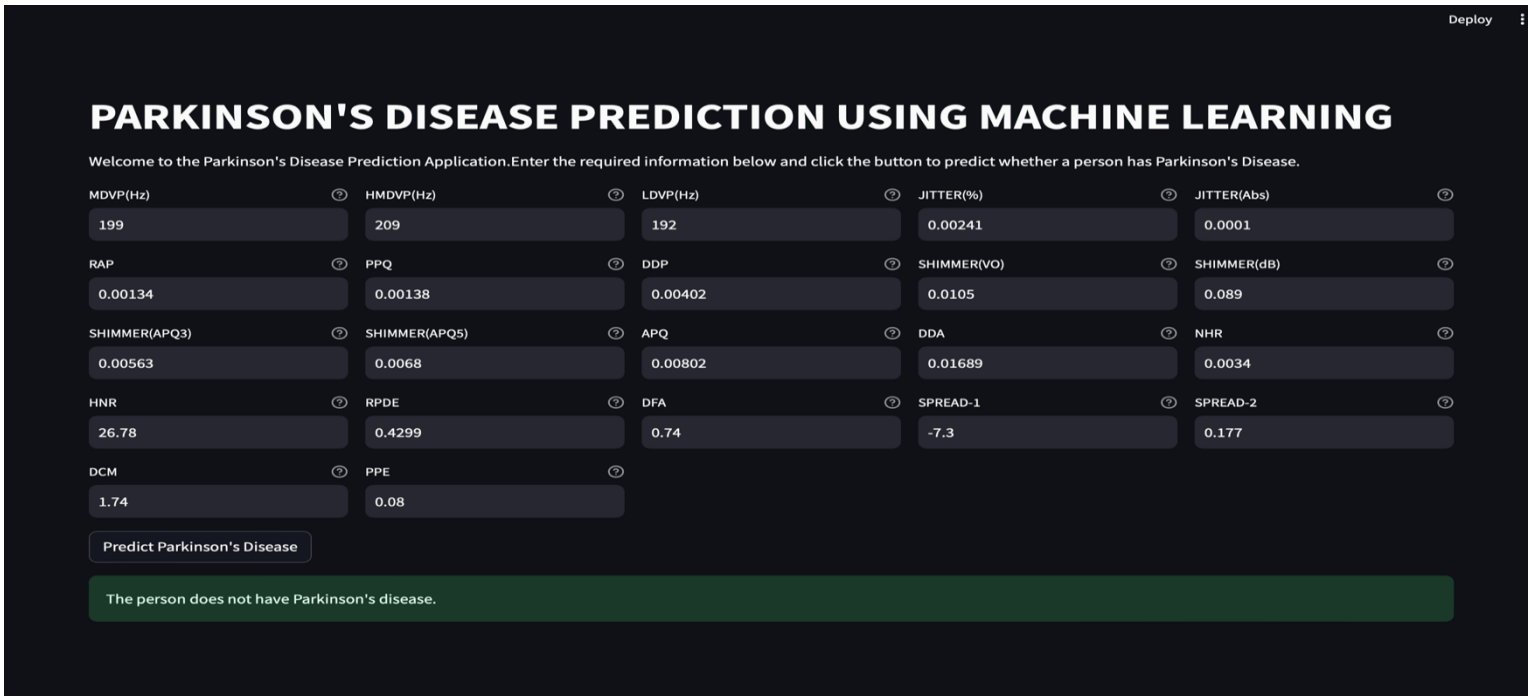


Figure 11:The model successfully predicting the person does not have parkinson's disease



VIII. Future Scope

A. Multi-Modal Data Integration:
Explore the integration of diverse data types, including genetic, imaging, and clinical data, to enhance the predictive power of models. Combining information from various sources can offer a deeper comprehension of PD and enhance diagnostic precision.

B. Longitudinal Studies and Temporal Analysis:
Conduct longitudinal studies to analyze disease progression over time. Temporal analysis using machine learning can capture dynamic changes in patient data, contributing to the early identification of subtle patterns indicative of PD onset.

C. Advanced Imaging Techniques:
Integrate various advanced imaging techniques such as functional MRI (fMRI) and positron emission tomography (PET) scans to capture more detailed neuroimaging characteristics. This can contribute to a more precise understanding of the neurodegenerative processes associated with PD.

D. Explainable AI (XAI) in Healthcare:
Incorporate Explainable AI methods to improve the clarity and understanding of machine learning models. This is crucial for gaining insights into the features contributing to predictions, fostering trust among healthcare practitioners and patients.

E. Real-Time Monitoring and Wearable Technology:
Develop real-time monitoring systems leveraging wearable technology for continuous data collection. This approach facilitates the creation of predictive models that can adapt to changes in patient health, enabling timely interventions and personalized care.

F. Patient-Centric Approach:
Shift towards a patient-centric approach by incorporating patient-reported outcomes and subjective well-being indicators. This can provide a more holistic view of the impact of PD on individuals and guide the development of patient-specific intervention strategies.

G. Ethical Considerations and Bias Mitigation:
Address ethical considerations related to data privacy, consent, and potential biases in machine learning models. Put in place strategies to guarantee equity and openness in the implementation of predictive models, especially in healthcare settings.

H. Collaborative Research Initiatives:
Encourage collaborative efforts among researchers, clinicians, and data scientists. Establishing interdisciplinary teams can foster the exchange of knowledge, methodologies, and datasets, accelerating progress in PD prediction research.

I. Integration with Electronic Health Records (EHR):
Explore seamless integration with Electronic Health Records to leverage the wealth of historical patient data. Integrating machine learning models into healthcare systems can facilitate real-world implementation and scalability.

J. External Validation and Generalization:
Conduct external validation studies across diverse populations to assess the generalizability of predictive models. Ensuring the robustness of models across different demographic groups enhances their clinical utility and relevance.

K. Public Health Applications:
Extend the application of PD prediction models to public health initiatives. This may involve identifying high-risk populations, implementing preventive measures, and contributing to the development of targeted public health campaigns.

This vision for the future highlights avenues to advance Parkinson's Disease prediction using machine learning, stressing the necessity for a thorough and evolving strategy to enhance diagnostic precision and patient results.

IX. Conclusion

Our investigation employed machine learning models, specifically Support Vector Machines (SVM), to predict the onset of Parkinson's Disease. Leveraging a comprehensive dataset comprising demographic, clinical, and biomarker information, our models demonstrated commendable accuracy in distinguishing between PD and non-PD cases. The predictive performance on both training and test datasets signifies the potential of machine learning as a valuable tool in early disease detection. The application of SVM allowed us to discern intricate patterns within the data, showcasing The importance of a characteristic or attribute. selection and model training. Notably, the inclusion of diverse features played a pivotal role in enhancing the robustness of our predictive models.

The effective utilization of machine learning models in predicting Parkinson's Disease holds substantial implications for the healthcare domain. Timely identification of Parkinson's Disease is pivotal for prompt interventions, the formulation of personalized treatment strategies, and enhanced patient outcomes. Leveraging machine learning for precise and efficient diagnosis contributes significantly to the broader realm of precision medicine.

Furthermore, our study emphasizes the potential integration of machine learning into clinical practice, augmenting the capabilities of healthcare professionals in diagnosing complex neurodegenerative disorders. The interpretability of our models provides a foundation for collaboration between AI systems and medical practitioners, fostering a synergistic approach to patient care.

Acknowledging the limitations of our study is imperative for a nuanced interpretation of the results. Challenges such as the lack of definitive biomarkers for PD, variations in data quality, and potential biases within the dataset underscore the complexity of predictive modeling in healthcare. Future research endeavors should address these limitations through the incorporation of diverse datasets, rigorous validation, and continuous model refinement.
Ethical considerations, including data privacy and the responsible deployment of machine learning in healthcare settings, remain paramount. Striking a balance between technological innovation and ethical standards is integral to the successful integration of predictive models into clinical workflows.

In conclusion, Our research represents progress in leveraging machine learning for predicting Parkinson's Disease, tapping into its potential. The amalgamation of data science and healthcare holds transformative promise, and while our study contributes to this narrative, it also underscores the need for ongoing innovation, collaboration, and ethical awareness.
As we navigate the complex landscape of neurodegenerative diseases, the integration of machine learning stands as a beacon of hope for early detection and improved patient care. Through the collective efforts of researchers, clinicians, and technologists, we envision a future where predictive models become integral to the proactive management of Parkinson's Disease, representing a significant change in healthcare approaches.

X. References

[1] Marras C, Beck JC, Bower JH, et al. Prevalence of Parkinson’s disease across North America. *NPJ Parkinsons Dis*. (2018.4.21)

[2]Chen, H. L., Huang, C. C., Yu, X. G., Xu, X., Sun, X., Wang, G., and Wang, S. J. (2013). An efficient diagnosis system for detection of Parkinson’s disease using fuzzy k-nearest neighbor approach. Expert systems with applications, 40(1), 263-271.

[3]Hariharan, M., Polat, K., and Sindhu, R. (2014). A new hybrid intelligent system for accurate detection of Parkinson's disease. Computer methods and programs in biomedicine, 113(3), 904-913.

[4]Åström, F., and Koker, R. (2011). A parallel neural network approach to prediction of Parkinson’s Disease. Expert systems with applications, 38(10), 12470-12474.

[5]Challa, K. N. R., Pagolu, V. S., Panda, G., and Majhi, B. (2016, October). An improved approach for prediction of Parkinson's disease using machine learning techniques. In Signal Processing, Communication, Power and Embedded System (SCOPES), 2016 International Conference on (pp. 1446-1451). IEEE.

[6]Agarwal, A., Chandrayan, S., and Sahu, S. S. (2016, March). Prediction of Parkinson's disease using speech signal with Extreme Learning Machine. In Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on (pp. 3776-3779). IEEE.

[7]Arora, S., Venkataraman, V., Zhan, A., Donohue, S., Biglan, K. M., Dorsey, E. R., and Little, M. A. (2015). Detecting and monitoring the symptoms of Parkinson's disease using smart phones: A pilot study. Parkinsonism & related disorders, 21(6), 650-653.

[8]Asgari, M., and Shafran, I. (2010, August). Predicting severity of Parkinson's disease from speech. In Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE (pp. 5201-5204). IEEE.

[9]Sriram, T. V., Rao, M. V., Narayana, G. S., and Kaladhar, D. S. V. G. K. (2016). ParkDiag: A Tool to Predict Parkinson Disease using Data Mining Techniques from Voice Data. International Journal of Engineering Trends and Technology (IJETT), 31(3), 136-140.

[10]Kim, J., Nasir, M., Gupta, R., Segbroeck, M. V., Bone, D., Black, M. P., and Narayanan, S. S. (2015). Automatic estimation of Parkinson's disease severity from diverse speech tasks. In Sixteenth Annual Conference of the International Speech Communication Association.

[11]Lewis, S. J. G., Foltynie, T., Blackwell, A. D., Robbins, T. W., Owen, A. M., & Barker, R. A. (2005). Heterogeneity of Parkinson's disease in the early clinical stages using a data driven approach. Journal of Neurology, Neurosurgery and Psychiatry, 76(3), 343-348.

[12]Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., and Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE transactions on biomedical engineering, 56(4), 1015-1022.

[13]Martínez-Martín, P., Rodríguez-Blázquez, C., Alvarez, M., Arakaki, T., Arillo, V. C., Chaná, P., and Serrano-Duenas, M. (2015). Parkinson's disease severity levels and MDS-Unified Parkinson's Disease Rating Scale. Parkinsonism and related disorders, 21(1), 50-54.

[14]Nilashi, M., Ibrahim, O., and Ahani, A. (2016). Accuracy improvement for predicting Parkinson's disease progression. Scientific reports, 6, 34181.

[15]Tsanas, A., Little, M. A., McSharry, P. E., and Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. IEEE transactions on Biomedical Engineering, 57(4), 884-893.

[16]Machine learning approaches to identify Parkinson's disease using voice signal features.Raya Alshammri , Ghaida Alharbi, Ebtisam Alharbi ,Ibrahim Almubark Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia(28/03/2023).

[17]Little, M.A., McSharry, P.E., Roberts, S.J. et al. (2007) "Exploiting Nonlinear Recurrence and Fractal Scaling Propertie s for Voice Disorder Detection" BioMed Eng OnLine 6 (23).

[18]Early detection of Parkinson's disease using machine learningAditi Govindu,(2023) Sushila Palwe School of SCET, MIT WPU, Pune 411038, India

[19]D. Barrejón, P. M. Olmos and A. Artés-Rodríguez, "Medical Data Wrangling With Sequential Variational Autoencoders," in IEEE Journalof Biomedical and Health Informatics, vol. 26, no. 6, pp. 2737-2745, June 2022, doi: 10.1109/JBHI.2021.3123839
.

[20]Tysnes,O.-B. ,and Storstein, A. (2017). Epidemiology of Parkinson's disease. J. Neural Trans. 124, 901–905. doi: 10.1007/s00702-017-1686-y

[21]Dorsey, E. R., Elbaz, A., Nichols, E., Abd-Allah, F., Abdelalim, A., Adsuar, J. C., et al. (2018). Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol. 17, 939–953. doi: 10.1016/S1474-4422(18)30295-3

[22]Drotár, P., Mekyska, J., Rektorová, I., Masarová, L., Smékal, Z., and Faundez-Zanuy, M. (2015). Decision support framework for Parkinson's disease based on novel handwriting markers. IEEE Trans. Neural Syst. Rehabil. Eng. 23, 508–516. doi: 10.1109/TNSRE.2014.2359997

[23]Pereira, C. R., Pereira, D. R., Rosa, G. H., Albuquerque, V. H. C., Weber, S. A. T., Hook, C., et al. (2018). Handwritten dynamics assessment through convolutional neural networks: An application to Parkinson's disease identification. Artif. Intell. Med. 87, 67–77. doi: 10.1016/j.artmed.2018.04.001

[24]Yang, M., Zheng, H., Wang, H., and McClean, S. (2009). "Feature selection and construction for the discrimination of neurodegenerative diseases based on gait analysis," in 2009 3rd International Conference on Pervasive Computing Technologies for Healthcare (London),1–7.doi: 10.4108/ICST.PERVASIVEHEALTH2009.6053

[25]Wahid, F., Begg, R. K., Hass, C. J., Halgamuge, S., and Ackland, D. C. (2015). Classification of Parkinson's disease gait using spatial-temporal gait features. IEEE J. Biomed. Health Inform. 19, 1794–1802. doi: 10.1109/JBHI.2015.2450232

[26]Pham, T. D., and Yan, H. (2018). Tensor decomposition of gait dynamics in Parkinson's disease. IEEE Trans. Bio-Med. Eng. 65, 1820–1827. doi: 10.1109/TBME.2017.2779884

[27]Cherubini, A., Morelli, M., Nisticó, R., Salsone, M., Arabia, G., Vasta, R., et al. (2014a). Magnetic resonance support vector machine discriminates between Parkinson disease and progressive supranuclear palsy. Move. Disord. 29, 266–269. doi: 10.1002/mds.25737

[28]Sakar, B. E., Isenkul, M. E., Sakar, C. O., Sertbas, A., Gurgen, F., Delil, S., et al. (2013). Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. IEEE J. Biomed. Health Inform. 17, 828–834. doi: 10.1109/JBHI.2013.2245674

[29]Lewitt, P. A., Li, J., Lu, M., Beach, T. G., Adler, C. H., Guo, L., et al. (2013). 3-hydroxykynurenine and other Parkinson's disease biomarkers discovered by metabolomic analysis. Move. Disord. 28, 1653–1660. doi: 10.1002/mds.25555

[30]Nuvoli, S., Spanu, A., Fravolini, M. L., Bianconi, F., Cascianelli, S., Madeddu, G., et al. (2019). [(123)I]Metaiodobenzylguanidine (MIBG) cardiac scintigraphy and automated classification techniques in Parkinsonian disorders. Mol. Imaging Biol. 22, 703–710. doi: 10.1007/s11307-019-01406-6

[31]Váradi, C., Nehéz, K., Hornyák, O., Viskolcz, B., and Bones, J. (2019). Serum N-glycosylation in Parkinson's disease: a novel approach for potential alterations. Molecules 24:2220. doi: 10.3390/molecules24122220

[32]Nunes, A., Silva, G., Duque, C., Januário, C., Santana, I., Ambrósio, A. F., et al. (2019). Retinal texture biomarkers may help to discriminate between Alzheimer's, Parkinson's, and healthy controls. PLoS One 14:e0218826. doi: 10.1371/journal.pone.0218826