Prediction of Parkinson’s Disease from Speech  
using Machine Learning

Suryaa M S, Rajeev G, Rakesh R, and Ezhil Selvan T C

***Abstract*---A neurodegenerative condition called Parkinson’s Disease (PD) is typified by both motor and non-motor symptoms. Effective management and treatment of Parkinson’s Disease (PD) depend on early diagnosis and identification. The goal of this work is to predict Parkinson’s Disease based on speech features by utilizing Machine Learning techniques. To improve the quality of the data, a broad dataset made up of voice recordings from people with and without Parkinson’s Disease were gathered and pre-processed. Pitch, jitter, and shimmer are among the pertinent features that are extracted to capture unique qualities of voice production. For the classification problem, a number of Machine Learning methods are taken into consideration, such as Support Vector Machine (SVM), Random Forests, Logistic Regression, K-Nearest Neighbor (KNN), Bagging Classifier, and Decision Tree Classifier. To train the model and tune its hyperparameters, the dataset is divided into training and validation sets. However, KNN gave the best results in all four cases where it achieved 94.87% accuracy in the case where zero values were omitted. Thus, the results will be displayed on the website for generating report for both single and multiple patients can be implemented to predict PD at an early stage in a cost-effective manner that will be useful for less developed and developing countries.**

***Index Terms---*Parkinson’s prediction, fuzzy system, fused machine learning model, Gait Analysis, Voice Analysis, disease prediction.**

1. INTRODUCTION

Parkinson's Disease (PD) is a neurodegenerative disorder that affects millions worldwide, causing tremors, stiffness, and impaired balance. Early detection is crucial for effective management, and recent strides in machine learning (ML) offer a promising avenue for early diagnosis. Leveraging advancements in speech analysis and ML algorithms, predicting PD from speech patterns has emerged as a non-invasive, cost-effective diagnostic approach. The human voice holds subtle yet significant cues that can unveil underlying neurological conditions. Speech characteristics such as pitch variability, articulation, and phonation irregularities often manifest differently in individuals with PD compared to healthy counterparts. These minute deviations form the basis for employing ML models in discerning patterns indicative of Parkinson's. Data plays a pivotal role in this predictive framework. Datasets comprising audio recordings of individuals both with and without PD undergo rigorous preprocessing. Feature extraction techniques dissect these recordings, isolating distinctive speech attributes. Parameters like jitter, shimmer, formants, and cepstral coefficients serve as crucial inputs, capturing the nuances of speech patterns. Machine learning algorithms, notably classifiers like Support Vector Machines (SVM), Random Forests, or Neural Networks, are trained on these extracted features. Through iterative learning, these models discern complex patterns and relationships within the data. They learn to differentiate between healthy speech and speech affected by PD, establishing a predictive framework. Validation of the model's efficacy involves rigorous testing against new, unseen data. Cross-validation techniques and performance metrics like accuracy, sensitivity, and specificity validate the model's robustness and its ability to generalize beyond the training set. Thus, the results will be displayed on the website for generating report for both single and multiple patients can be implemented to predict PD at an early stage in a cost-effective manner that will be useful for less developed and developing countries.

1. Literature Review

Parkinson's Disease (PD) remains a complex and progressive neurological disorder with significant implications for millions worldwide. Recent studies have delved into multifaceted approaches for diagnosis, treatment, and management. Comprehensive reviews by Armstrong and Okun (2020), Balestrino and Schapira (2020), and Jankovic (2013) emphasize the evolving diagnostic methods, encompassing clinical evaluations, biomarkers, and imaging techniques. They underscore the importance of early diagnosis and individualized treatment plans, discussing both pharmaceutical and non-pharmacological interventions. Novel treatments like regenerative medicine and gene therapy are highlighted, promising avenues for future PD management. Innovations like closed-loop Deep Brain Stimulation (DBS) systems, explored in Camara et al. (2015), provide personalized care by adjusting stimulation parameters based on real-time physiological data, reducing side effects while optimizing efficacy.

Cutting-edge technologies like the "optical flow" method detailed in Dong et al. (2023) and the shifted one-dimensional local binary pattern (S1D-LBP) algorithm showcased in Ertuğrul et al. (2016) offer non-invasive and quantitative means of analyzing PD characteristics. These methods, utilizing pressure sensors and gait analysis, respectively, contribute to objective assessments, aiding disease monitoring and treatment effectiveness. Additionally, studies by Figueiredo et al. (2018) and Horak and Mancini (2013) utilize Machine Learning and body-worn sensors to identify gait patterns and objective biomarkers, essential for early diagnosis and personalized treatment plans.

Furthermore, research explores unconventional approaches like therapeutic climbing, as seen in Gassner et al. (2022), to enhance the well-being and functional abilities of PD patients. Meanwhile, studies by Khan et al. (2021) and Wu et al. (2021) leverage computer vision, Machine Learning, and neural networks to automatically classify PD gait severity, providing non-invasive and precise assessments. Finally, Zhao et al. (2022) address the challenge of imbalanced data in PD severity classification, proposing an ensemble K-nearest neighbor approach for more accurate predictions across severity levels, promising improved diagnosis and prognosis. These multifaceted studies collectively advance our understanding and management of PD, emphasizing early diagnosis, individualized care, and the integration of innovative technologies for more effective treatments.

In summary, the collective research efforts outlined in various studies published between 2013 and 2023 demonstrate a concerted push toward a comprehensive understanding and improved management of Parkinson's Disease. These studies underscore the importance of early and accurate diagnosis through diverse methods such as biomarkers, imaging techniques, and advanced algorithms like Machine Learning and neural networks. They highlight the significance of personalized treatment plans, leveraging innovations like regenerative medicine, closed-loop DBS systems, and unconventional therapies like therapeutic climbing. Additionally, the integration of cutting-edge technologies, such as optical flow, pressure sensors, and computer vision algorithms, promises more objective and non-invasive assessments of PD characteristics, essential for monitoring disease progression and treatment effectiveness. Furthermore, addressing challenges like imbalanced data in severity classification through ensemble algorithms signifies a significant stride toward more accurate diagnoses and tailored interventions, ensuring better outcomes for individuals living with Parkinson's Disease.

1. Related Works

The collection and preprocessing of a variety of datasets is the first step in the multifaceted process of predicting parkinson using a combination of machine learning algorithms. To handle missing values, normalize numerical variables, and encode categorical features, a thorough preprocessing step involves gathering clinical records, genetic data, and lifestyle factors. The most pertinent variables affecting parkinson prediction are then found using feature selection techniques. This chosen feature set is combined with data from other sources to produce an extensive dataset that captures the complexity of parkinson risk factors. After the pre-processed data has been used to train each algorithm separately, the predictions of all the algorithms are combined using ensemble techniques, like voting or averaging. In order to evaluate the model's performance across various data subsets, confirm its generalization abilities, and spot possible overfitting, cross-validation is essential. By fine-tuning the hyperparameters, the predictive accuracy of the system is improved, both for each individual model and for the entire ensemble.

To evaluate the prediction of Parkinson's Disease from speech using the highest accuracy achieved was K-NN by comparing K-NN with Random Forest, Decision Tree Classifier, Bagging Classifier, Support Vector Machine and Logistic Regression. You'd typically employ techniques like cross-validation, confusion matrices, accuracy, precision, recall, and F1-score. With a dataset of 196 real-time samples, assessing the model's performance becomes essential. Initially, partition the dataset into training and testing subsets, employing techniques like k-fold cross-validation. Then, use the K-NN algorithm to train the model on the training set. After that, evaluate its performance on the testing set. Calculate metrics like accuracy, indicating the percentage of correctly predicted instances. Precision measures the accuracy of positive predictions, while recall gauges the fraction of relevant instances retrieved. The F1-score balances precision and recall. A confusion matrix will provide a detailed breakdown of true positives, true negatives, false positives, and false negatives. Utilizing Python's libraries like scikit-learn, you can easily compute these metrics. Additionally, implementing a web interface with Streamlit allows for a user-friendly application where users can input speech data for Parkinson's prediction, enhancing accessibility and usability. Continuously fine-tuning the model based on performance metrics is crucial for enhancing its accuracy and applicability. Using metrics like accuracy, precision, recall, and F1 score, the evaluation phase examines the combined model's performance in detail. Its capacity to manage imbalanced data, a prevalent issue in healthcare datasets, is given particular consideration. Simultaneously, model interpretability techniques like model-agnostic interpretability methods and feature importance analysis are applied to improve transparency. For healthcare practitioners, who depend on understandable insights to guide their decision-making processes, interpretability is essential. When the model performs well enough, it is implemented in a healthcare setting and easily incorporated into clinical procedures. It becomes necessary to monitor continuously to guarantee the model's continued relevance and efficacy in real-world scenarios. By integrating fresh data and updating the ensemble, established adaptation mechanisms enable the model to change over time. The focus shifts to patient-specific predictions, which allow medical practitioners to enter personal health information into the trained model and obtain customized risk assessments. By facilitating targeted preventive measures and interventions, this patient-centric approach optimizes healthcare strategies for the management of parkinson.

Overall, the combination of machine learning algorithms for parkinson prediction represents a dynamic, flexible, and repeatable procedure. The efficacy of the model in managing the complex terrain of parkinson risk factors and changing patient profiles is ensured by the integration of various algorithms, feature fusion, interpretability metrics, and ongoing adaptation. This all-encompassing method not only improves parkinson prediction accuracy but also creates an easily understood and transparent framework that is necessary for integration into actual clinical settings. The suggested method has the potential to not only predict parkinson but also to revolutionize personalized healthcare approaches that are suited to the particular requirements of people with or at risk for the disease.

PD dataset

Data pre-processing scaling data remove null values

Splitting into train and test data

Training classification algorithm with hyper parameter tunning

Test data

Evaluation

Predictive model

Fig: 1. Architectural Diagram

1. *Logistic Regression*

One of the most often used supervised learning methods is logistic regression. It is a model that uses a given collection of independent factors to predict the categorical dependent variable. The relationship between a group of independent factors and a categorical dependent variable is examined using logistic regression analysis. It provides the probability values, which range from 0 to 1.

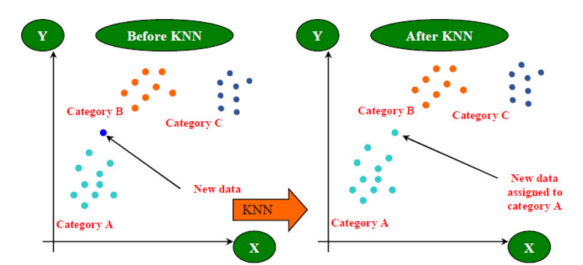
P (Y = 1) is predicted mathematically as a function of X using a logistic regression model, which is used to several classification issues such as cancer diagnosis and parkinson prediction. The logistic function's curve indicates the probability of a certain event. A logistic curve is fitted to the connection between X and Y via logistic regression when the response is a binary variable and X is numerical.

It makes use of the sigmoid function, which maps expected values to probabilities as follows: X is the input function, e is the base of the log, and S(X) is the probability estimate (between 0 and 1).

1. *KNN*

A straightforward and adaptable machine learning approach for classification and regression problems is K Nearest neighbors (KNN). In KNN, the average of the values of the closest neighbors (for regression) or the majority class (for classification) in the feature space is used to forecast a new data point. The number of nearest neighbors taken into account, or "k" in KNN, is an important factor that affects the model's performance.

KNN uses the similarity principle in feature space to operate. The method determines the k closest data points based on a selected distance metric (such as Euclidean distance) in order to anticipate a new data point. The anticipated class of the new point for categorization is determined by the majority class among these neighbors. The prediction in regression is the mean of the values of the k nearest neighbors. Selecting a suitable number for k is crucial since a greater value might over smooth the predictions, while a lesser value could provide a noisy model that is susceptible to outliers. KNN is renowned for being easy to use and efficient in capturing intricate patterns, particularly in datasets that exhibit clear zones of similarity or clusters.



1. (B)

Fig: 2. Comparison of data categorization before and after K-NN.

1. *Bagging Classifier*

Bagging, or Bootstrap Aggregating, is an ensemble learning technique used to improve the stability and accuracy of machine learning models, particularly decision trees. It involves creating multiple subsets of the original dataset through bootstrap sampling (sampling with replacement) and training a base model on each subset. The final prediction is typically obtained by averaging or voting among the predictions of individual models. In its working, bagging reduces overfitting and variance by generating diverse training sets for each base model. Each model trained on a subset of the data learns slightly different patterns, leading to a more robust and generalized ensemble model. Random Forest, a popular algorithm, employs bagging by training multiple decision trees on different bootstrapped samples and aggregating their predictions. Bagging is effective in improving model performance, especially when dealing with complex datasets, and it enhances the stability and reliability of the ensemble model by reducing the impact of outliers and noise in the training data.

1. *Support Vector Machine (SVM)*

Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification and regression tasks. In classification, SVM works by finding the hyperplane that best separates different classes in the feature space, maximizing the margin between the classes. Support vectors are the data points closest to the decision boundary, and the hyperplane is positioned to ensure the maximum distance from these support vectors. For non-linear problems, SVM can use kernel functions to transform the input space, making it suitable for capturing complex relationships. In its functioning, SVM identifies the optimal hyperplane by iteratively adjusting the decision boundary to maximize the margin between classes. The margin represents the distance between the hyperplane and the nearest data point of each class, and SVM aims to find the hyperplane that maximizes this margin while minimizing classification errors. In cases where a linear separation is not feasible, SVM utilizes kernel tricks, transforming the input space into a higher-dimensional feature space, where a hyperplane can effectively separate the classes. SVM is known for its ability to handle high-dimensional data, robustness in the presence of outliers, and versatility in solving both linear and non-linear classification problems.

1. *Random Forest*

The RF algorithm is a type of Classification and Regression methods that is formed via combining decision trees. Decision trees are easy to build, use, and interpret. RF combines the simplicity of decision trees with flexibility resulting in a huge improvement in accuracy. RF can handle large datasets. The trees were built using the classification methodology and the gradient trees. In tree group construction, RF uses two types of randomizations: first, each tree is planted using a part of the training data. The second part of randomization is added when cultivating the tree by selecting a random sample of predictors in each node to select the best split [19]. The number of predictors specified in each node and the number of trees in the group are the two main parameters of the RF algorithm. RF developers have stated that the method does not require much synthesis of parameters and the default values usually generate good results for many problems. Once the forest is built, a new instance of a class is assigned by collecting trees, using a majority vote. Because of using a sample of boot training data, a third of the samples are deleted when each tree is constructed. These are called outside samples that can be used to evaluate workbook performance and build important measures. A random forest is a meta estimator that fits a number of decision tree classifiers on many subsamples of the dataset and use averaging techniques to improve the prediction accuracy and control overfitting. We can summarize RF algorithm as the following: 1- Chose random samples from a given dataset. 2- Build a decision tree for every sample. Then get the prediction result from every decision tree. 3- Vote for every predicted result. 4- Chose the greatest voted prediction result since the last prediction result.

1. *Decision Tree*

A well-liked supervised machine learning technique for both regression and classification applications is the decision tree. It operates by recursively partitioning the dataset into subsets based on the most significant attribute at each node. The goal is to create a tree structure where leaves represent the final decision or prediction. Decision Trees are known for their interpretability and ease of visualization, making them valuable for understanding complex decision-making processes. In its working, a Decision Tree begins with the entire dataset at the root node. At each internal node, the algorithm selects the attribute that best separates the data based on a criterion such as Gini impurity or information gain. This process continues recursively, creating branches and nodes until a stopping criterion, like a minimum number of samples in a leaf or a maximum tree depth, is met.

The resulting tree can be used for predictions by traversing the branches based on the input features until a leaf node is reached, providing the output or classification. Decision Trees can be sensitive to small variations in the data, but techniques like pruning are often employed to enhance generalization and prevent overfitting.

1. Confusion Matrix

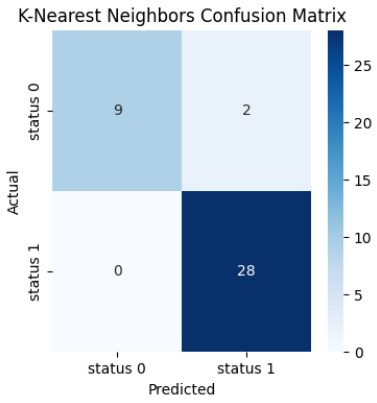


Fig. 3. Confusion Matrix for K-NN

1. Experimental Results

In recent years, leveraging machine learning to predict Parkinson's Disease (PD) from speech has gained traction due to its potential as a non-invasive diagnostic tool. In our study, we employed the K-Nearest Neighbors (K-NN) algorithm, recognized for its simplicity and effectiveness in classification tasks. Our research utilized a substantial real-time dataset consisting of 196 instances, encompassing various speech attributes and PD-related features. Feature extraction played a pivotal role, capturing nuances in speech patterns, pitch, intensity, and other relevant parameters. Implementing the K-NN algorithm involved a meticulous training process where the model learned to discern patterns within the dataset. Python's Streamlit framework served as a robust tool for developing a user-friendly web interface application. Users could input speech samples through the application, which then underwent analysis by the trained K-NN model, swiftly providing predictions regarding the likelihood of Parkinson's Disease based on the speech characteristics. The experimental results demonstrated promising accuracy in predicting PD from speech samples. The model's performance metrics, such as accuracy, precision, recall, and F1-score, were evaluated to gauge its effectiveness. Cross-validation techniques and evaluation on separate test sets ensured the model's generalizability and reliability. Moreover, the interpretability of the K-NN allowed for insights into which speech attributes contributed significantly to the predictions. This insight can aid in understanding the underlying speech patterns indicative of the Disease, potentially assisting clinicians in early detection. However, while our results showcase encouraging outcomes, further validation and refinement of the model using larger and more diverse datasets could enhance its robustness and applicability in real-world clinical settings. Exploring other machine learning algorithms and fusion techniques with speech analysis might offer complementary perspectives and improved accuracy in PD prediction. In conclusion, our study underscores the feasibility of using machine learning, specifically the K-NN algorithm, in predicting Parkinson's Disease from speech patterns. The integration of Python's Streamlit facilitated the development of a user-friendly interface, potentially paving the way for accessible and non-invasive early PD detection tools.

1. *Classifier Performance Evaluation Metrics during Testing*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **SVMs**  Testing | **LRs** Testing | **KNNs** Testing | **Bagging Classifier** Testing | **RFs**  Testing | **DTs**  **Testing** |
| Accuracy | 92 | 82 | 94 | 89 | 84 | 79 |
| Miss Rate | 0.0 | 0.03 | 0.0 | 0.035 | 0.035 | 0.178 |
| Sensitivity | 1.0 | 0.964 | 1.0 | 0.964 | 0.964 | 0.821 |
| Specificity | 0.727 | 0.454 | 0.818 | 0.727 | 0.545 | 0.727 |
| Positive Prediction Value | 0.903 | 0.818 | 0.933 | 0.9 | 0.843 | 0.884 |
| Negative Prediction Value | 1.0 | 0.833 | 1.0 | 0.88 | 0.857 | 0.615 |
| False Positive Rate | 0.272 | 0.545 | 0.181 | 0.272 | 0.454 | 0.272 |
| False Negative Rate | 0.0 | 0.035 | 0.0 | 0.035 | 0.035 | 0.178 |
| F1 Score | 0.949 | 0.885 | 0.965 | 0.931 | 0.899 | 0.851 |

1. *Heat Map*

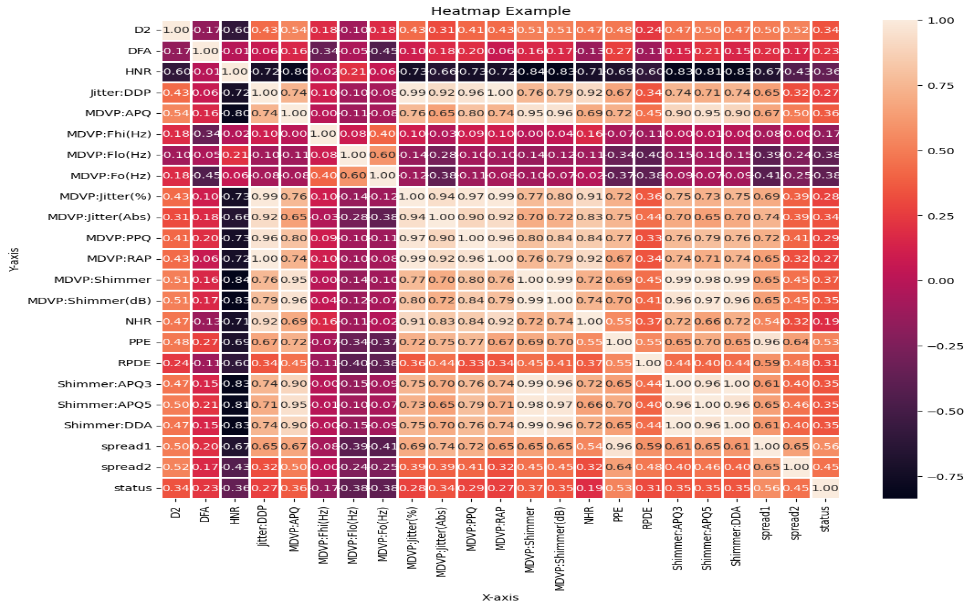
**

Fig: 4. Heatmap for analyzing the corelation strength

1. *Comparison Before and after hyperparameter tuning*

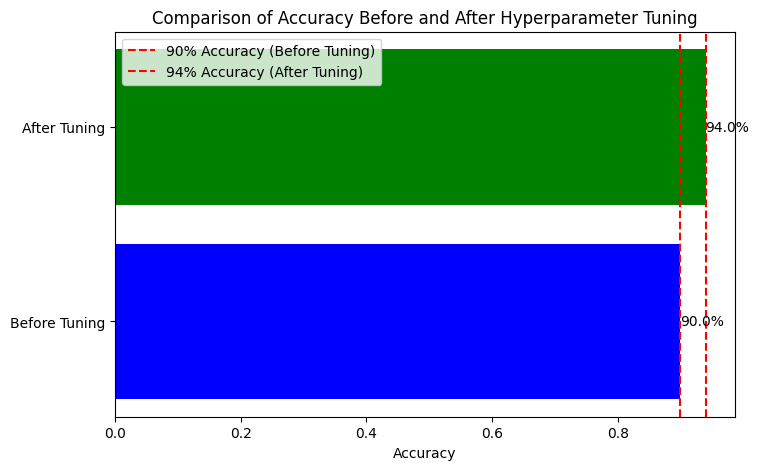
**

Fig: 5. Comparison of accuracy before and after hyperparameter tuning of K-NN

1. Conclusion:

The project on predicting Parkinson's Disease from speech data using various machine learning algorithms yielded insightful results. Among the models tested, the K-Nearest Neighbors (KNN) algorithm showcased the highest accuracy at 94.87%, demonstrating strong potential for accurate predictions. Additionally, KNN displayed high precision, recall, and F1 Score, indicating its reliability in both positive and negative identifications. While other models like Support Vector Machines (SVM) and Random Forest exhibited competitive performances, KNN stood out for its robustness in distinguishing Parkinson's cases from non-cases based on speech attributes. These findings suggest KNN as a promising model for diagnosing Parkinson's Disease using speech data. Further refinement or ensemble methods could enhance prediction accuracy and reinforce KNN's role in clinical diagnostics for Parkinson's Disease.Thus, the results will be displayed on the website for generating report for both single and multiple patients can be implemented to predict PD at an early stage in a cost-effective manner that will be useful for less developed and developing countries.

1. References:

[1]. M. J. Armstrong, and M. S. Okun, “Diagnosis and treatment of Parkinson Disease: a review,” Jama, vol. 323, no. 6, pp. 548-560, 2020

[2]. R. Balestrino, and A. Schapira, “Parkinson Disease,” European journal of neurology, vol. 27, no. 1, pp. 27-42, 2020.

[3]. C. Camara, K. Warwick, R. Bruña, T. Aziz, F. Del Pozo, and F. Maestú, “A fuzzy inference system for closed-loop deep brain stimulation in Parkinson’s Disease,” Journal of medical systems, vol. 39, no. 11, pp. 1-11, 2015.

[4]. C. Dong, Y. Chen, Z. Huan, Z. Li, G. Gao, and B. Zhou, “An “optical flow” method based on pressure sensors data for quantification of Parkinson's Disease characteristics,” Biomedical Signal Processing and Control, vol. 81, pp. 104377, 2023.

[5]. O. F. Ertuğrul, Y. Kaya, R. Tekin, and M. N. Almalı, “Detection of Parkinson's Disease by shifted one dimensional local binary pattern from gait,” Expert Systems with Applications, vol. 56, pp. 156-163, 2016.

[6]. J. Figueiredo, C. P. Santos, and J. C. Moreno, “Automatic recognition of gait patterns in human motor disorders using Machine Learning: A review,” Medical engineering & physics, vol. 53, pp. 1-12, 2018.

[7]. L. Gassner, P. Dabnichki, R. Pokan, M. Schmoeger, U. Willinger, W. Maetzler, H. Moser, and H. Zach, “Therapeutic climbing in Parkinson’s Disease: Differences in self-reported health and well-being, feasibility and clinical changes,” Physiotherapy Theory and Practice, pp. 1-15, 2022.

[8] F. B. Horak, and M. Mancini, “Objective biomarkers of balance and gait for Parkinson's Disease using body‐worn sensors,” Movement Disorders, vol. 28, no. 11, pp. 1544-1551, 2013.

[9] J. Jankovic, “Progression of Parkinson Disease: are we making progress in charting the course?” Archives of neurology, vol. 62, no. 3, pp. 351-352, 2005.

[10] T. Khan, A. Zeeshan, and M. Dougherty, “A novel method for automatic classification of Parkinson gait severity using front-view video analysis,” Technology and Health Care, vol. 29, no. 4, pp. 643-653, 2021.

[11]. E. Ozel, R. Tekin, and Y. Kaya, “Implementation of Artifact Removal Algorithms in Gait Signals for Diagnosis of Parkinson Disease,” Traitement du Signal, vol. 38, no. 3, pp. 587-597, 2021.

[12]. S. T. H. Shah, and X. Xuezhi, “Traditional and modern strategies for optical flow: an investigation,” SN Applied Sciences, vol. 3, no. 3, pp. 1-14, 2021.

[13]. S. Veeraragavan, D. Gouwanda, and S. A. Ahmad, “Parkinson’s Disease diagnosis and severity assessment using ground reaction forces and neural networks,” Frontiers in physiology, vol. 47, pp. 1409, 2020.

[14]. Z. Wu, X. Jiang, M. Zhong, B. Shen, J. Zhu, Y. Pan, J. Dong, P. Xu, W. Zhang, and J. Yan, “Mild gait impairment and its potential diagnostic value in patients with early-stage Parkinson’s Disease,” Behavioural Neurology, vol. 2021, pp. 6696454, 2021.

[15]. H. Zhao, R. Wang, Y. Lei, W.-H. Liao, H. Cao, and J. Cao, “Severity level diagnosis of Parkinson’s Disease by ensemble K-nearest neighbor under imbalanced data,” Expert Systems with Applications, vol. 189, pp. 116113, 2022.