**Early Detection Of Parkinson’s Disease Using Machine Learning**

|  |  |  |
| --- | --- | --- |
| Mridul Madnani  *20BDS0191*  *Vellore Institute of Technology, Vellore*  [mridul.madnani2020@vitstudent.ac.in](mailto:mridul.madnani2020@vitstudent.ac.in) | Srijan Singh Somvanshi  *20BDS0381*  *Vellore Institute of Technology, Vellore*  [srijansingh.somvanshi2020@vitstudent.ac.in](mailto:srijansingh.somvanshi2020@vitstudent.ac.in) | Devansh Bajpai  *20BCE0807*  *Vellore Institute of Technology, Vellore*  [devansh.bajpai2020@vitstudent.ac.in](mailto:devansh.bajpai2020@vitstudent.ac.in) |
|  |  |  |

# **1.INTRODUCTION**

Parkinson's disease (PD) is a chronic and progressive neurological disorder that affects millions of people worldwide, particularly those over the age of 60. The disease is characterized by the loss of dopamine-producing neurons in the brain, leading to motor symptoms such as tremors, rigidity, bradykinesia (slowness of movement), and postural instability. Apart from motor symptoms, PD can also cause non-motor symptoms such as cognitive impairment, depression, and sleep disturbances, significantly impacting patients' quality of life. Early diagnosis of Parkinson's disease is critical for timely intervention and effective management. However, diagnosing PD in its early stages remains challenging due to the subtle onset of symptoms and the overlap with other neurological conditions. Current diagnostic methods primarily rely on clinical assessments, which are subjective and may not detect early-stage PD accurately. The objective of this project is to develop an advanced machine learning-based system for the early identification of Parkinson's disease using vocal biomarkers and a stacking ensemble technique. By leveraging machine learning algorithms and data analysis techniques, the system aims to enhance diagnostic accuracy and facilitate timely intervention, ultimately improving patient outcomes.

# **2.LITERATURE SURVEY**

The survey on existing systems for Parkinson's disease (PD) diagnosis reveals several insights into current approaches and their limitations. Existing systems predominantly rely on traditional machine learning models such as logistic regression, support vector machines (SVM), decision trees, and random forests for PD detection. These systems often face challenges related to feature selection, data heterogeneity, and class imbalance, which can impact diagnostic accuracy.

Feature selection in traditional PD diagnostic systems is typically based on manual identification of potential biomarkers, such as tremor severity or gait abnormalities, which may not capture the full complexity of PD symptoms. Furthermore, the reliance on individual classifiers limits the ability to leverage diverse information from different data sources effectively.

Moreover, existing systems often struggle with handling data heterogeneity arising from variations in data formats, patient demographics, and disease progression stages. This heterogeneity can introduce bias and affect the generalizability of machine learning models across different populations. Class imbalance, another common issue in PD diagnosis, occurs when the number of positive (PD) cases is significantly lower than negative (non-PD) cases in the dataset. This imbalance can lead to biased model predictions and affect the overall performance of the diagnostic system.

A diagram of a disease

Description automatically generated

Figure 1: Pathological Progression vs Time

The survey also highlights the limited adoption of ensemble learning techniques in PD diagnosis. Ensemble methods, such as stacking, offer the potential to combine predictions from multiple base models to improve diagnostic accuracy and robustness. However, their application in PD diagnostics remains underexplored.

In summary, the survey underscores the need for an advanced diagnostic system that addresses the limitations of existing approaches by:

- Implementing sophisticated feature selection techniques to identify comprehensive biomarkers of PD.

- Leveraging ensemble learning methods to integrate diverse information and enhance diagnostic accuracy.

- Handling data heterogeneity and class imbalance effectively to improve model generalizability and reliability.

By bridging these gaps, the proposed PD diagnostic system aims to advance the state-of-the-art in machine learning-based healthcare diagnostics and contribute to early and accurate detection of Parkinson's disease.

# **3.PROBLEM STATEMENT**

Parkinson's disease (PD) poses a significant challenge due to its progressive nature and varied symptomatology, necessitating early detection for optimal management. This project addresses the development of an advanced machine learning system using a stacking ensemble technique to detect PD in its early stages. Key challenges include selecting relevant features from diverse data sources, integrating multiple algorithms for improved accuracy, handling class imbalances in datasets, and rigorously evaluating model performance. By overcoming these hurdles, the system aims to provide a robust diagnostic tool capable of effectively distinguishing PD from non-PD cases, thereby contributing to early detection and better management of Parkinson's disease.

# **4.OBJECTIVE**

The objectives of the project break down the aim into specific, measurable, and achievable tasks or goals. These objectives serve as the roadmap for the project, guiding the implementation and evaluation process. Each objective should be clear, actionable, and aligned with the overall aim of the project. For example, objectives may include collecting and analyzing relevant data, developing predictive algorithms, evaluating the performance of the model, and validating the results.

* Develop a machine learning model for the early detection of Parkinson’s disease.
* Improve Diagnostic Precision: Enhance accuracy through effective feature selection and engineering.
* Facilitate Timely Intervention: Enable prompt medical management by integrating the model into healthcare systems.

# **5.METHODOLOGY**

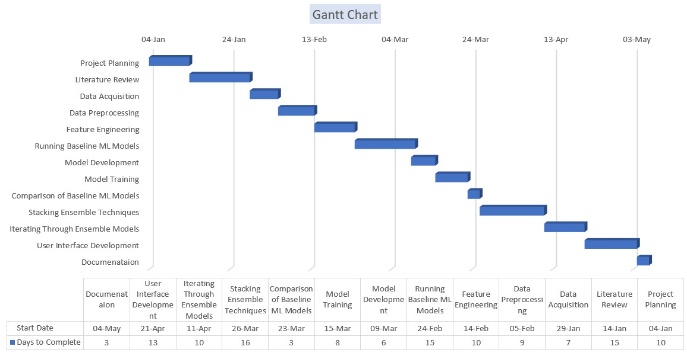


Figure 2: Gantt Chart depicting workflow timeline

**Module - 1: Data Acquisition and Preprocessing**

In this module, our primary focus is on gathering raw data pertinent to Parkinson's disease (PD) and laying the groundwork for subsequent analysis. We commence by sourcing datasets from a multitude of channels including clinical databases, ongoing research studies, and publicly available repositories, ensuring a comprehensive coverage of relevant information. Once acquired, our attention shifts to data cleaning and harmonization, where we meticulously address inconsistencies, errors, and missing values within the dataset. Through rigorous quality control measures, we implement robust data validation procedures to uphold the accuracy and reliability of the information, laying a solid foundation for subsequent analysis and model development.

**Module - 2: Preprocessing and Feature Engineering**

In this module, our primary focus is on gathering raw data associated with Parkinson's disease (PD) and readying it for analysis. This involves several key steps, starting with data collection from diverse sources like clinical databases, research studies, and public repositories. Subsequently, we perform data cleaning and harmonization to eliminate inconsistencies, errors, and missing values, ensuring the dataset's integrity. Quality control measures are then implemented to validate the data, guaranteeing its accuracy and reliability throughout the analysis process.

**Module - 3: Running Baseline ML Models**

This module encompasses training and evaluating baseline machine learning models using preprocessed data for Parkinson's disease diagnosis. It includes implementing various algorithms such as logistic regression, K-nearest neighbors (KNN), Naive Bayes, random forest, decision tree, support vector machine (SVM), and XGBoost. For each algorithm, we optimize hyperparameters and assess model performance using metrics like accuracy, precision, recall, and F1-score. These models range from traditional ones like logistic regression and decision trees to more complex ensemble methods like random forest and XGBoost, offering a comprehensive approach to predictive modeling for Parkinson's disease detection.

**Module - 4: Model Development and Training**

This module focuses on meticulously selecting and fine-tuning machine learning algorithms to build accurate predictive models for Parkinson's disease diagnosis. Initially, a comprehensive evaluation of diverse algorithms such as logistic regression, K-nearest neighbors, naive Bayes, random forest, decision tree, support vector machine, and XGBoost is conducted, considering performance metrics like accuracy and interpretability. Following this, hyperparameters are optimized to enhance model performance and generalization, striking a balance between complexity and efficacy. The preprocessed dataset is then split into training, validation, and test sets for model training, utilizing cross-validation techniques to assess performance across various data subsets and mitigate overfitting. Iterative optimization iterates through refining model architectures, hyperparameters, and training strategies based on validation results, experimenting with different features and techniques to enhance model robustness and generalization.

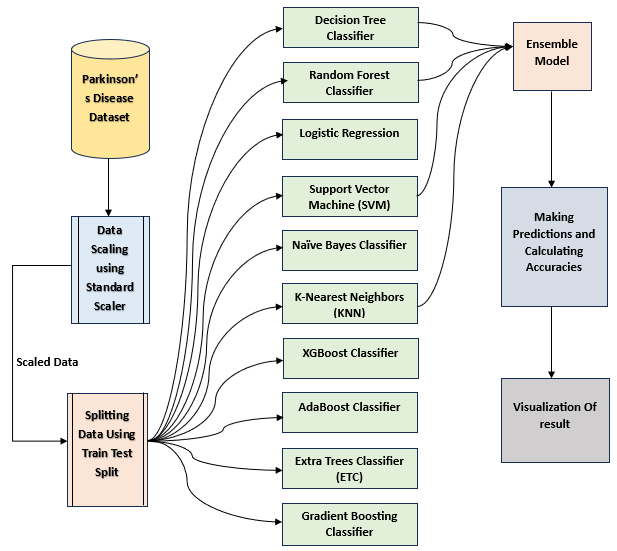
****

Figure 3: Data Flow Diagram

**Module - 5: Comparison and Stacking Ensemble Techniques**

This module focuses on comparing the performance of individual baseline models and implementing ensemble techniques to enhance predictive accuracy. It involves evaluating baseline models using metrics like accuracy, precision, recall, and F1-score, considering factors such as complexity and interpretability. Stacking ensemble methods are then utilized to combine predictions from diverse base models, training a meta-learner to improve overall performance. Ensemble model performance is assessed using appropriate metrics and validation techniques, comparing it with individual baseline models. The ensemble model's interpretability and decision-making process are analyzed, visualizing decision boundaries to gain insights into its behavior.

**Module - 6: Validation and Evaluation**

This module focuses on evaluating the performance of developed models against baseline algorithms, using validation techniques to assess accuracy, precision, recall, and F1-score. It includes comparing model metrics with baseline algorithms, identifying strengths and weaknesses, and selecting effective models for clinical deployment. Performance is visualized using ROC curves and confusion matrices, analyzing trade-offs between metrics. Interpretability is assessed, enhancing model transparency by explaining predictions. Model selection involves refining promising models based on evaluation results, adjusting parameters, and incorporating domain knowledge for optimization.

**Module - 7: Building a User Interface**

This module is dedicated to crafting a user-friendly web application interface that seamlessly integrates our optimized ensemble model for the early detection of Parkinson's disease. Key features include designing intuitive input fields for users to enter patient data, ensuring clear guidance and error handling for accurate submissions. The interface triggers the ensemble model to generate predictions based on user input, prominently displaying outcomes. Visual representations enhance understanding, with charts or diagrams illustrating prediction probabilities. The design is responsive across devices, with interactive elements allowing real-time adjustments. Leveraging HTML/CSS, JavaScript, and frontend frameworks like React.js or Vue.js ensures dynamic UI components and interactivity, optimizing user experience and accessibility

# **6.RESULTS**



Table 1: Evaluation results summarizing model performance using key metrics.

The developed ensemble stacking model, which integrates Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) classifiers, exhibited exceptional accuracy, achieving approximately 98% accuracy in the early detection of Parkinson's disease. This impressive result underscores the robustness and effectiveness of the ensemble learning approach, which leverages the complementary strengths of diverse individual classifiers. By combining the predictive abilities of multiple models, the ensemble method outperforms standalone baseline classifiers like logistic regression or decision trees, offering superior accuracy and robustness in identifying PD cases.

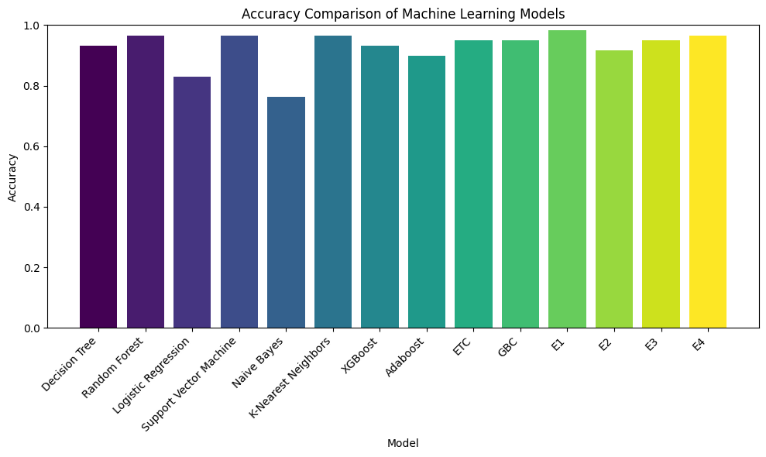


Figure 4: Visualization showing the accuracy of different ML models & their Ensembles.

# **7.CONCLUSION**

The project aims to develop an advanced machine learning-based system for the early detection of Parkinson's disease (PD) using a stacking ensemble technique, addressing gaps identified in existing PD diagnostic systems. By leveraging ensemble learning, the system integrates feature selection and multiple machine learning algorithms to enhance diagnostic accuracy and effectiveness.

Existing PD diagnostic systems often rely on individual classifiers, which may struggle with data heterogeneity and class imbalance. The proposed system addresses these limitations by employing a stacking ensemble approach, which combines predictions from diverse base models to improve overall performance.

The project's objectives include implementing feature selection techniques to identify discriminative PD-related features and developing a stacked ensemble model integrating various machine learning algorithms. Through rigorous evaluation and comparison with baseline classifiers, the effectiveness of the proposed ensemble approach is demonstrated.

Ensemble learning offers several advantages over traditional single-model approaches. By combining predictions from multiple base models (e.g., logistic regression, decision trees, support vector machines), ensemble methods can capture different aspects of the data and reduce overfitting. This diversity and aggregation of predictions lead to improved accuracy and robustness in PD detection compared to individual classifiers.

In real-world usage, the developed PD diagnostic system can provide clinicians with more reliable and interpretable diagnostic results, facilitating early intervention and personalized treatment plans. The system's effectiveness in handling data heterogeneity, class imbalance, and feature selection contributes to its practical utility and potential impact in clinical settings.

By rectifying gaps in existing PD diagnostic systems and demonstrating the superiority of ensemble learning over baseline models, this project advances the state-of-the-art in machine learning-based healthcare diagnostics. The proposed system holds promise for improving patient outcomes through early and accurate detection of Parkinson's disease.

# **8.REFERENCES**

[1] P. R. Magesh, R. D. Myloth, and R. J. Tom, “An Explainable Machine Learning Model for Early Detection of Parkinson’s Disease using LIME on DaTSCAN Imagery,” Computers in Biology and Medicine, vol. 126, p. 104041, Nov. 2020, doi: [https://doi.org/10.1016/j.compbiomed.2020.104041.](https://doi.org/10.1016/j.compbiomed.2020.104041)

[2] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," in IEEE Access, vol. 5, pp. 8869-8879, 2017, doi: 10.1109/ACCESS.2017.2694446.

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

[4] R. Alshammri, G. Alharbi, E. Alharbi, and I. Almubark, “Machine learning approaches to identify Parkinson’s disease using voice signal features,” Frontiers in Artificial Intelligence, vol. 6, p. 1084001, Mar. 2023, doi: <https://doi.org/10.3389/frai.2023.1084001.>

[5] S. K. Biswas, A. Nath Boruah, R. Saha, R. S. Raj, M. Chakraborty, and M. Bordoloi, “Early detection of Parkinson disease using stacking ensemble method,” Computer Methods in Biomechanics and Biomedical Engineering, pp. 1–13, May 2022, doi: [https://doi.org/10.1080/10255842.2022.2072683.](https://doi.org/10.1080/10255842.2022.2072683)

[6] A. Govindu and S. Palwe, “Early detection of Parkinson’s disease using machine learning,” Procedia Computer Science, vol. 218, pp. 249–261, 2023, doi: [https://doi.org/10.1016/j.procs.2023.01.007.](https://doi.org/10.1016/j.procs.2023.01.007)

[7] Karapinar Senturk, Z. Early Diagnosis of Parkinson’s Disease Using Machine Learning Algorithms. Medical Hypotheses 2020, 138, 109603. [https://doi.org/10.1016/j.mehy.2020.109603.](https://doi.org/10.1016/j.mehy.2020.109603)

[8] A. Govindu and S. Palwe, “Early detection of Parkinson’s disease using machine learning,” Procedia Computer Science, vol. 218, pp. 249–261, 2023, doi: <https://doi.org/10.1016/j.procs.2023.01.007.>

[9] D. D. Joshi, H. H. Joshi, B. Y. Panchal, P. Goel and A. Ganatra, "A Parkinson Disease Classification Using Stacking Ensemble Machine Learning Methodology," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 1335-1341, doi: <https://ieeexplore.ieee.org/document/9823509>

[10] T. Velmurugan and J. Dhinakaran, “A Novel Ensemble Stacking Learning Algorithm for Parkinson’s Disease Prediction,” Mathematical Problems in Engineering, vol. 2022, pp. 1–10, Jul. 2022, doi: <https://doi.org/10.1155/2022/9209656.>

[11] H. V. Nguyen and H. Byeon, “Prediction of Parkinson’s Disease Depression Using LIME-Based Stacking Ensemble Model,” Mathematics, vol. 11, no. 3, p. 708, Jan. 2023, doi: <https://doi.org/10.3390/math11030708.>

[12] K. M. Alalayah, Ebrahim Mohammed Senan, H. F. Atlam, Ibrahim Abdulrab Ahmed, and A. Shatnawi, “Automatic and Early Detection of Parkinson’s Disease by Analyzing Acoustic Signals Using Classification Algorithms Based on Recursive Feature Elimination Method,” vol. 13, no. 11, pp. 1924–1924, May 2023, doi: <https://doi.org/10.3390/diagnostics13111924.>

[13] S. A. Doumari, K. Berahmand, and M. J. Ebadi, “Early and High-Accuracy Diagnosis of Parkinson’s Disease: Outcomes of a New Model,” Computational and Mathematical Methods in Medicine, vol. 2023, p. 1493676, 2023, doi: <https://doi.org/10.1155/2023/1493676.>

[14] Shelke, Maya & Ranjan, Nihar & Mate, Gitanjali. (2022). Detection of Parkinson's Disease using Machine Learning Algorithms and Handwriting Analysis. Journal of Data Mining and Management. 8. 10.46610/JoDMM.2023.v08i01.004.

[15] Johann Faouzi, Olivier Colliot, and Jean-Christophe Corvol, “Machine Learning for Parkinson’s Disease and Related Disorders,” Neuromethods, pp. 847–877, Jan. 2023, doi: <https://doi.org/10.1007/978-1-0716-3195-9_26.>