International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Advanced Parkinson's Prediction System Using Random Forest Algorithm and Feature Engineering

A. Alagar[1], U. Balaji[2], S.S. Barath[3], K. Gokul Krishnan[4]

[1]*Assistant Professor, Department of Computer Science and Engineering, K.L.N. College of Engineering*

[2,3,4] *Final Year Students, Department of Computer Science and Engineering, K.L.N. College of Engineering*

Abstract - This system offers an enhanced Parkinson's Disease Prediction Model utilizing machine learning for prompt diagnosis. By utilizing health and lifestyle information, it utilizes Random Forest for categorization, assessing its precision against K-Nearest Neighbors (KNN). The model goes through data preprocessing, feature engineering, and assessment to improve prediction accuracy. An interface built on Streamlit enables users to enter patient information and obtain real-time predictions along with confidence scores. Furthermore, analyzing feature importance assists in recognizing crucial contributing elements, aiding in medical decision-making. By prioritizing real-time forecasting, model interpretability, and scalability, this system supports early identification and improved disease management. Upcoming enhancements involve speech and walking assessment, incorporation of genetic information, and advanced deep learning frameworks for improved precision and usability .

***Key Words*:** Machine Learning, Parkinson's Disease Prediction, Random Forest, K-Nearest Neighbors (KNN), Data Preprocessing, Feature Engineering, Model Evaluation, Confidence Score

1. INTRODUCTION

Parkinson's disease is a progressive neurological condition that impacts movement, making early identification vital for proper management. This system utilizes machine learning methods to estimate the probability of Parkinson's disease using health and lifestyle information. Employing Random Forest for classification and assessing its performance against K-Nearest Neighbors (KNN) guarantees high accuracy and dependability. Data preprocessing and feature engineering improve model performance, while an intuitive Streamlit interface enables smooth user interaction for real-time predictions accompanied by confidence scores. Furthermore, analyzing feature importance reveals significant elements affecting the disease, supporting more effective medical decision-making. This system, created for scalability and real-time forecasting, offers an innovative and understandable method for predicting Parkinson's disease. Upcoming enhancements seek to combine speech and gait analysis, genetic information, and deep learning methods to improve accuracy and functionality.

2. OBJECTIVE

The objective of this system is to create a precise and effective machine learning-driven method for anticipating Parkinson's disease, facilitating early diagnosis and improved disease management. By examining a mix of medical, lifestyle, and demographic information, the system employs sophisticated classification methods, chiefly Random Forest, to differentiate between Parkinson's-positive and Parkinson's-negative cases. The model's performance is assessed in comparison to K-Nearest Neighbors (KNN) to guarantee optimal outcomes. To improve accessibility, the system features an intuitive interface powered by Streamlit, allowing healthcare professionals and researchers to enter patient data and obtain real-time predictions with confidence scores. Furthermore, it emphasizes interpretability by integrating feature importance analysis, enabling medical professionals to pinpoint key factors that impact disease prediction. This clarity supports clinical decision-making, providing crucial insights into patient health. Future improvements seek to enhance the model by incorporating deep learning methodologies, genetic markers, and real-time monitoring for more thorough and dependable predictions.

3. LITERATURE SURVEY

3.1. Machine Learning in Parkinson's Disease Diagnosis:

Traditional diagnostic methods for Parkinson's disease depend on clinical assessments, neurological evaluations, and imaging technologies like MRI and PET scans. These approaches are frequently lengthy, costly, and necessitate specialized medical knowledge, resulting

in postponed diagnoses. To address these obstacles, researchers have investigated machine learning (ML) methods for early identification by examining health and lifestyle factors. Different ML models, including Support Vector Machines (SVM), Decision Trees, and Neural Networks, have shown encouraging results in recognizing Parkinson's symptoms based on patient information. These models analyze vocal characteristics, hand tremors, gait patterns, and demographic information to enhance prediction accuracy. Nevertheless, data preprocessing, feature selection, and class imbalance continue to pose considerable challenges in optimizing ML models for practical medical use. Improving these models through effective data engineering and hybrid algorithms can enhance their dependability and flexibility. As AI-enhanced healthcare progresses, refining ML techniques can facilitate quicker, more precise Parkinson's diagnoses, supporting early intervention and treatment.

3.2. Feature Importance and Its Role in Prediction Accuracy:

Accurate prediction of diseases relies on determining the most pertinent features that aid in diagnosis. Investigations into feature selection methods, including Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE), underscore their success in optimizing models for improved accuracy and efficiency. In the context of Parkinson's disease prediction, vocal tremors, speech anomalies, hand movements, and lifestyle habits act as fundamental indicators. By employing these feature selection strategies, machine learning models can concentrate on the most valuable data points, minimizing computational demands while upholding predictive accuracy. Explainable AI approaches, like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations), further bolster interpretability by shedding light on the features that most significantly impact predictions. This clarity is vital in medical contexts, where grasping risk factors is critical for making clinical decisions. The combination of feature selection and explainability enhances the dependability of models, allowing healthcare practitioners to diagnose Parkinson's disease with increased accuracy and assurance.

4. EXISTING SYSTEM

The Parkinson's Disease Prediction System employs machine learning to evaluate voice metrics for the early identification of Parkinson's disease. It implements group-wise scaling to mitigate biases associated with age and biological sex, enhancing accuracy by 9. 5% compared to traditional methods. The model reaches 82% accuracy on previously unseen data, guaranteeing its dependability in practical applications. Shapley Additive Explanations (SHAP) values are utilized to elucidate model choices, offering insights into the elements affecting classification. Results demonstrate that consistent voice patterns with frequent pauses are strongly linked to Parkinson's disease, while healthy individuals show greater variability in voiced segments, elevated pitch variation, and increased spectral flux. These insights correspond with established medical research, bolstering the model's efficacy in disease detection. By incorporating advanced feature engineering and interpretability methods, this system improves diagnostic accuracy and aids healthcare professionals in early intervention. Future enhancements may encompass deep learning integration and the addition of further biometric features for increased accuracy.

5. PROPOSED SYSTEM

Achieving an accuracy of around 93%, the Parkinson's Disease Prediction System employs a Random Forest model, surpassing K-Nearest Neighbors (KNN) in classification effectiveness. By examining health and lifestyle information, the system facilitates early identification, guaranteeing accurate and reliable forecasts. A web-based interface and command-line tools allow for seamless engagement, enabling users to train models and create predictions with ease. Confidence scores are provided with every prediction, improving result transparency and assisting healthcare providers in making well-informed choices. To enhance accuracy, data preprocessing methods are utilized, ensuring consistency and dependability in input data. Furthermore, visualization tools like confusion matrices and classification reports offer more profound insights into the model's effectiveness. Future developments are intended to incorporate deep learning models, real-time patient oversight, and supplementary biometric attributes, further enhancing the system's accuracy and flexibility in medical diagnostics. These enhancements will strengthen the system's function in the early detection of Parkinson's disease and healthcare decision-making.

6. ARCHITECTURE DIAGRAM

This system enhances the prediction of Parkinson's disease by combining data collection, preprocessing, machine learning, and user interaction. Initially, medical and demographic data are gathered and

cleansed through methods like addressing missing values, feature scaling, and normalization to guarantee dependability. Subsequently, the refined data is utilized to train Random Forest and K-Nearest Neighbors (KNN) models, with accuracy and performance metrics determining the most suitable choice. The developed models are saved and accessed through a command-line interface or an easy-to-use Streamlit web application for real-time predictions. To boost model performance, feature selection methods such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) enhance inputs while hyperparameter tuning assures optimal outcomes. Regularization methods avoid overfitting, and assessment metrics like precision, recall, and F1-score improve classification dependability. Built for scalability, the system accommodates cloud deployment, with upcoming improvements integrating deep learning and real-time patient monitoring to advance early diagnosis and treatment approaches.

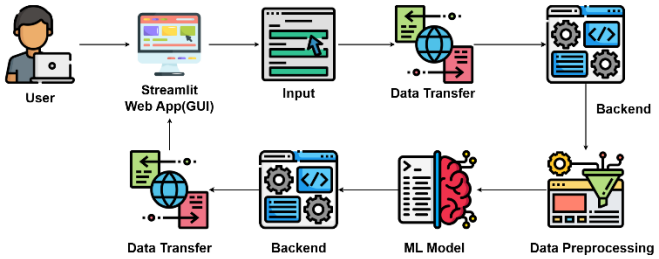

Figure 1: Architecture Diagram

7. SYSTEM OVERVIEW

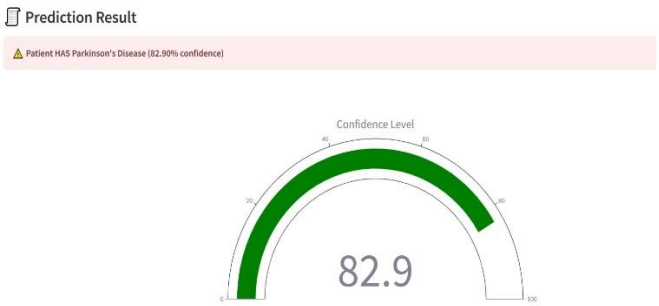

Figure 2: Prediction Result (Positive)

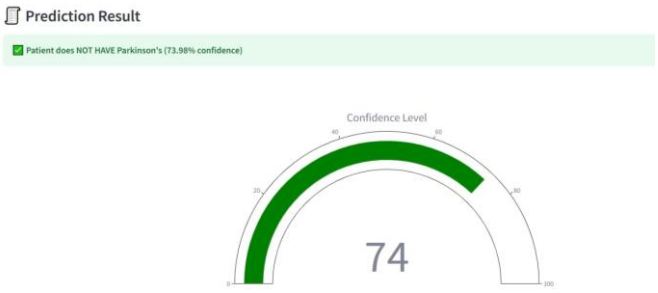

Figure 3: Prediction Result (Negative)

7.1. Data Collection and Preprocessing:

The system starts by collecting comprehensive patient information, which encompasses medical history, lifestyle choices, and various health metrics. Ensuring the quality of data is essential for precise predictions, thus the assembled dataset undergoes rigorous preprocessing. This process includes addressing missing values, detecting and correcting anomalies, and removing redundant or unrelated features that could impair model performance. Normalization and feature scaling methods are utilized to maintain uniformity in data distribution, enhancing the model's learning effectiveness. Feature engineering is applied to derive significant insights by converting raw data into more pertinent attributes that aid in the classification of Parkinson's disease. Methods such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) assist in pinpointing the most significant variables, lowering computational complexity while preserving predictive strength. The enhanced dataset is subsequently divided into training and testing portions to assess model performance, guaranteeing that the system provides dependable and strong predictions for early disease identification.

7.2. Machine Learning Model Training:

To create an efficient prediction system for Parkinson's disease, two machine learning models—Random Forest and K-Nearest Neighbors (KNN)—are trained and assessed. Random Forest, a powerful ensemble learning technique, has shown exceptional classification results, attaining an accuracy of 93%. KNN, while more straightforward, acts as a comparative benchmark to confirm the effectiveness of Random Forest. The training of the model involves thorough hyperparameter tuning utilizing techniques like Grid Search and Random Search, refining parameters such as tree depth and the quantity of estimators for Random Forest. The dataset is segmented into training and validation sets to avoid overfitting and improve generalizability. Once training is complete, the models

are stored effectively, enabling immediate application when new patient data is introduced. The trained models are subjected to additional validation using performance metrics like precision, recall, F1-score, and confusion matrices to guarantee their capability to accurately distinguish between Parkinson's-positive and Parkinson's-negative cases with minimal false positives and false negatives.

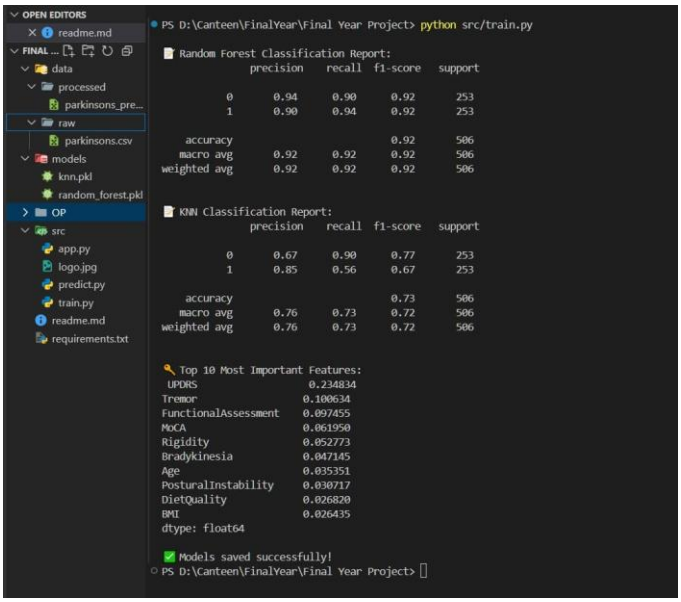

Figure 4: Model Training

7.3. Prediction and Result Interpretation:

The prediction phase permits users to input patient information via a command-line interface or a user-friendly web application created with Streamlit. The system analyzes the input data, inputting it into the trained Random Forest model to produce predictions accompanied by a confidence score. This score signifies the level of certainty the model has in its classification, assisting healthcare professionals in evaluating reliability. To maintain interpretability, the system uses Shapley Additive Explanations (SHAP) values, which delineate the contribution of each feature to the final decision. Significant influencing factors, such as voice tremors, variations in speech, and movement irregularities, are emphasized, enabling medical experts to corroborate predictions with clinical observations. The results are presented in an accessible format, complete with graphical representations, confusion matrices, and rankings of feature importance, making it simpler for healthcare providers to make well-informed choices about patient diagnosis, treatment strategies, and possible follow-up evaluations.

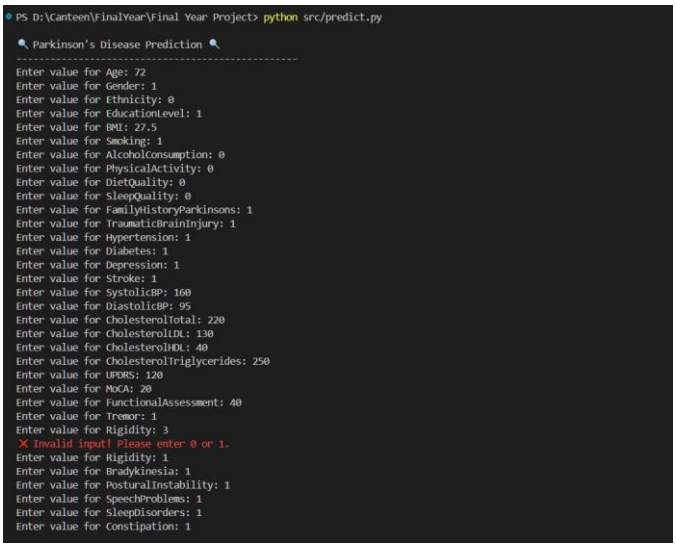

Figure 5 Model Evaluation Result

7.4. User Interface and Deployment:

To improve accessibility and usability, a web-based interface is created using Streamlit, offering a smooth user experience for healthcare professionals and researchers. The interactive platform enables users to enter patient data and obtain instant predictions, removing the necessity for intricate technical skills. Made for scalability, the system can be implemented on cloud platforms, guaranteeing remote access and connection with hospital databases. Real-time prediction features provide immediate feedback, assisting doctors and caregivers in making prompt decisions. The interface additionally includes data visualization capabilities, showcasing prediction results with clear explanations, increasing trust and reliability. Security protocols, such as encrypted data transmission and authentication measures, ensure patient data privacy and adherence to medical standards. Prospective enhancements involve integrating deep learning models for more accurate predictions and adding real-time monitoring functions using wearable devices, further improving the system's relevance in contemporary healthcare environments.

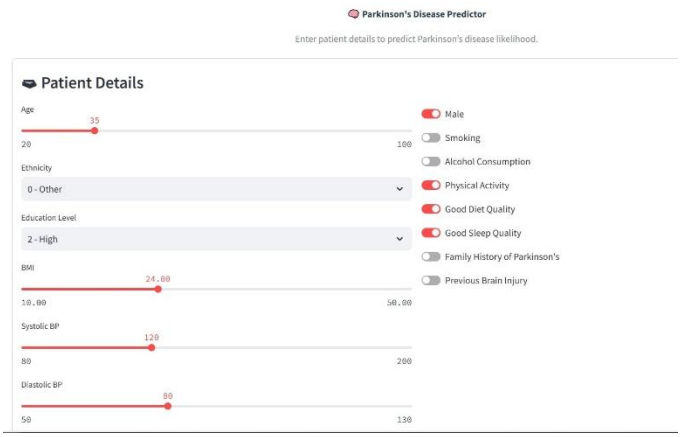

Figure 6: User Interface

8. FUTURE ENHANCEMENT

The Parkinson's Disease Prediction System can be significantly improved by integrating advanced machine learning models such as deep learning architectures like LSTMs and CNNs, which can enhance accuracy and adaptability. Additionally, implementing federated learning will enable decentralized training, ensuring patient data privacy while continuously improving the model. Another major enhancement involves IoT and wearable device integration, allowing real-time monitoring of patient symptoms through smartwatches and health sensors. This continuous data stream can refine predictions, provide early warnings, and track disease progression. Furthermore, personalized insights will be introduced, where the system learns from patient history and adapts to offer customized recommendations.

To enhance accessibility and scalability, the system can be deployed on cloud platforms such as AWS, GCP, or Azure, enabling multi-user access for medical professionals and researchers. A more interactive user interface with data visualizations and a mobile-friendly design will improve usability. Telemedicine integration will allow doctors to review predictions remotely and sync with Electronic Health Records (EHRs) for seamless medical data access. Security measures such as encryption and compliance with HIPAA and GDPR will be strengthened. Additionally, multi-language support and ccessibility features will improve inclusivity. Lastly, automated model updates and collaboration with medical institutions will help expand datasets, ensuring continuous improvement and research advancements.

9. CONCLUSION

The Parkinson's Disease Prediction System utilizes machine learning to improve early detection and patient care. By analyzing health and lifestyle information, the system applies Random Forest, achieving an accuracy of 93%, which surpasses KNN in classification dependability. Predictions are accompanied by confidence scores, ensuring clarity in the decision-making process. A straightforward Streamlit interface facilitates smooth interaction for researchers and healthcare experts. The system analyzes input data in real-time, delivering immediate diagnostic insights. Planned future updates include deep learning integration and real-time patient monitoring to enhance accuracy and user experience. By progressing AI-driven diagnostics, the system aids in delivering more accurate, accessible, and effective detection and management of Parkinson's disease.

REFERENCES

[1]. N. Momeni, S. Whitling, and A. Jakobsson, "Interpretable Parkinson's Disease Detection Using Group-Wise Scaling", IEEE Access, vol. 13, pp. 29147-29161, 2025.

[2]. N. Momeni, S. Whitling, and A. Jakobsson, "Detecting Parkinson's disease using voice recordings from mobile devices", Proc. 32nd Eur. Signal Process. Conf. (EUSIPCO), pp. 1516-1520, Aug. 2024.

[3]. C. Botelho, A. Abad, T. Schultz, and I. Trancoso, "Speech as a Biomarker for Disease Detection", IEEE Access, vol. 12, pp. 184487-184508, 2024.

[4]. C. R. Dhivyaa, K. Nithya, and S. Anbukkarasi, "Enhancing Parkinson's Disease Detection and Diagnosis: A Survey of Integrative Approaches Across Diverse Modalities", IEEE Access, vol. 12, pp. 158999-159024, 2024.

[5]. A. Siderowf et al., "Assessment of heterogeneity among participants in the Parkinson's progression markers initiative cohort using α-synuclein seed amplification: A cross-sectional study", Lancet Neurol., vol. 22, no. 5, pp. 407-417, May 2023.

[6]. M. K. Reddy and P. Alku, "Exemplar-Based Sparse Representations for Detection of Parkinson's Disease From Speech", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 1386-1396, 2023.

[7]. C. Dong, Y. Chen, Z. Huan, Z. Li, B. Zhou, and Y. Liu, "Static-Dynamic Temporal Networks for Parkinson's Disease Detection and Severity Prediction", IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 2205-2213, 2023.

[8]. S. Turaev et al., "Review and Analysis of Patients' Body Language From an Artificial Intelligence Perspective", IEEE Access, vol. 11, pp. 62140-62173, 2023.

[9]. S. J. J. Jui, R. C. Deo, P. D. Barua, A. Devi, J. Soar, and U. R. Acharya, "Application of Entropy for Automated Detection of Neurological Disorders With Electroencephalogram Signals: A Review of the Last Decade (2012–2022)", IEEE Access, vol. 11, pp. 71905-71924, 2023.

[10]. H. Khachnaoui, B. Chikhaoui, N. Khlifa, and R. Mabrouk, "Enhanced Parkinson's Disease Diagnosis Through Convolutional Neural Network Models Applied to SPECT DaTSCAN Images", IEEE Access, vol. 11, pp. 91157-91172, 2023.