International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Identification of Gene Variant Associated with Parkinson's Disease Using Genomic Databases

Ms. Karpagavalli K[1], Sona P[2], Yoga Lakshmi N[3], Yuvasri E[4]

[1]Assistant Professor, Department of CSE, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India
[2,3,4]Student, Department of CSE, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India

Abstract:
Parkinson's Disease (PD) is a progressive neurodegenerative disorder that severely impacts motor and non-motor functions. This study aims to identify gene variants associated with PD using genomic databases and bioinformatics tools. Leveraging databases such as NCBI and Ensembl, and applying machine learning algorithms like Logistic Regression and Random Forest, the project classifies gene variants as pathogenic or benign. Feature selection is based on parameters like clinical significance, variant frequency, and pathogenicity scores (SIFT, PolyPhen, CADD). This work provides a foundation for improving early diagnosis, personalizing treatments, and understanding the genetic basis of PD.

Keywords: Parkinson's Disease, Genomic Databases, Gene Variants, Machine Learning, Logistic Regression, Random Forest

INTRODUCTION

Parkinson's Disease affects dopamine-producing neurons in the substantia nigra, causing symptoms such as tremors, rigidity, and bradykinesia. Traditional diagnostic methods are limited in early detection and accuracy. As genetic factors are known to influence PD, the project focuses on detecting gene variants associated with the disease using genomic analysis. By integrating bioinformatics tools and machine learning, the project proposes a novel, data-driven method for identifying critical mutations, enabling early intervention and contributing to personalized medicine.

METHODOLOGY

The methodology for identifying gene variants associated with Parkinson's Disease (PD) involves a comprehensive and multi-phased approach that integrates genomic data analysis with machine learning-based classification techniques. The system begins by sourcing raw genomic variant data from well-established biological repositories such as the National Center for Biotechnology Information (NCBI), Ensembl, and UniProt. These platforms provide extensive datasets that include variant information, clinical annotations, and predictive scores which are essential for distinguishing between benign and pathogenic mutations. Data collection focuses on genes and variants that have been previously associated with neurodegenerative conditions, particularly Parkinson's Disease.

Once the data is gathered, the preprocessing stage plays a vital role in preparing it for machine learning models. This phase involves cleaning the dataset by removing duplicates and handling missing or

inconsistent values. Standard imputation methods are used when appropriate, and outliers are identified using statistical techniques such as the interquartile range or Z-score. Numerical features such as gene expression levels or pathogenicity scores are normalized to ensure they fall within a consistent scale, which improves model performance. Categorical variables such as clinical significance (e.g., "benign", "likely pathogenic") are encoded into numerical formats using label encoding or one-hot encoding to be machine-readable. The preprocessing stage ensures the data is not only clean and structured but also optimally formatted for model training and evaluation.

Feature selection is the next critical step in the methodology. Due to the high dimensionality of genomic data, not all features contribute meaningfully to the prediction process. To enhance model efficiency and reduce overfitting, statistical techniques like chi-square tests, correlation analysis, and mutual information scores are applied to determine which features most significantly impact the classification outcome. Additionally, machine learning-based feature selection methods such as Recursive Feature Elimination (RFE) and feature importance scores from ensemble models like Random Forest are utilized. This step ensures that the model focuses on key predictors such as SIFT, PolyPhen, and CADD scores, along with variant frequency and clinical annotation tags.

The core of the methodology lies in the training and evaluation of the machine learning models. Two supervised learning algorithms—Logistic Regression and Random Forest—are employed due to their effectiveness in binary classification problems. Logistic Regression is chosen for its simplicity, interpretability, and suitability in medical datasets, while Random Forest is leveraged for its ability to handle complex relationships and deliver higher accuracy through ensemble learning. The dataset is split into training and testing subsets, typically in a 70-30 ratio. During training, hyperparameters are optimized using grid search and cross-validation techniques to ensure robustness and prevent overfitting. Both models learn to differentiate between pathogenic and non-pathogenic variants based on the training data.

After training, the models are evaluated using a variety of performance metrics. Accuracy, precision, recall, and F1-score are calculated to assess the overall performance. The Area Under the ROC Curve (AUC-ROC) is used to visualize the trade-off between sensitivity and specificity. A confusion matrix is also generated to illustrate the number of true positives, false positives, true negatives, and false negatives. To increase model interpretability, especially in the healthcare context, tools like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) are employed. These tools help explain how each feature contributed to a specific prediction, making the model's decision process transparent for clinicians and researchers.

The final stage of the methodology involves deploying the trained models for real-time prediction. In this phase, new genetic data samples are input into the system, and the models classify each variant as either likely pathogenic or benign. This step bridges the gap between research and practical implementation, allowing for quick risk assessment in clinical or research settings. The integration of the model into a cloud-based development environment such as Google Colab ensures easy accessibility, scalability, and collaborative capabilities.

Overall, the methodology combines bioinformatics, data science, and machine learning to build a scalable and efficient system for the identification of Parkinson's-related gene variants. It enables more accurate early diagnosis, supports personalized medicine, and paves the way for further genetic studies in neurodegenerative diseases

PROPOSED SYSTEM

The proposed system aims to accurately identify and classify gene variants associated with Parkinson's Disease (PD) using an integrative approach that combines genomic data mining with machine learning techniques. The system is designed to extract and process genomic variant data from trusted databases such as NCBI, Ensembl, and UniProt, which provide comprehensive information on known gene mutations, protein structures, and associated pathogenicity scores. Once the data is collected, it undergoes a series of preprocessing steps, including cleaning, normalization, and feature encoding, to ensure that it is ready for computational analysis. Key features such as SIFT, PolyPhen, and CADD scores are extracted, as they provide valuable insight into the potential impact of each variant on protein function. These features are then fed into machine learning classifiers—Logistic Regression and Random Forest—which have been selected for their strong performance in binary classification tasks. The system trains on annotated genetic data to distinguish between benign and pathogenic variants. Once trained, the models are capable of evaluating new genomic samples, offering real-time predictions on the likelihood of Parkinson's-related gene involvement. Hosted in a cloud-based environment like Google Colab, the system is highly scalable, accessible, and optimized for collaboration, making it suitable for both academic research and clinical support applications. Through this approach, the proposed system provides a robust and scalable solution for early identification of PD-associated gene variants, aiding in genetic research, personalized medicine, and potential early intervention strategies.

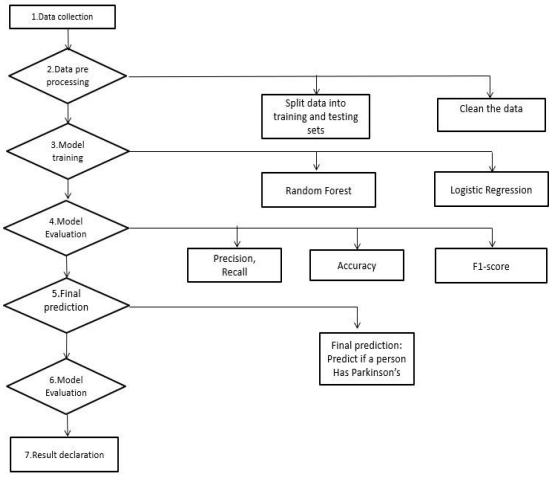


Fig.1.Architecture of Proposed System

EXISTING SYSTEM

The current methods for diagnosing Parkinson's Disease (PD) largely depend on clinical observations, patient history, and neurological examinations such as the Unified Parkinson's Disease Rating Scale (UPDRS). Imaging technologies like MRI, PET scans, and transcranial sonography have also been employed to aid diagnosis. However, these techniques often detect changes only after the disease has significantly progressed, limiting their usefulness for early-stage detection. Moreover, they are costly, time-consuming, and sometimes vary in accuracy depending on patient demographics and clinical interpretation. Traditional computational methods for studying genetic variants primarily rely on threshold-based detection or static rule-based models. These methods analyze gene data by setting fixed criteria—such as mutation frequency or deviation from known genetic markers—to flag abnormalities. While simple to implement, they are rigid, lack adaptability to new patterns, and struggle with complex

datasets, often leading to high rates of false positives and missed early indicators of PD.
Additionally, while tools like SIFT, PolyPhen, and CADD offer valuable individual insights into gene variant pathogenicity, they are often used in isolation without being incorporated into a cohesive predictive framework. Conventional genetic studies have mostly focused on single mutations or small sets of genes rather than building comprehensive models that can analyze multiple features simultaneously. There is also minimal integration of machine learning, which restricts the capacity to recognize hidden patterns and interactions within large genomic datasets. As the volume of genetic data grows and the understanding of PD's genetic complexity deepens, these existing systems are becoming increasingly insufficient. This highlights a critical need for more dynamic, intelligent systems that can automate variant classification, adapt to new genetic findings, and enhance the accuracy of early Parkinson's diagnosis using a data-driven, scalable approach.

SYSTEM DESIGN

The process begins with Genomic Data Input, where variant details such as SNP identifiers, gene names, clinical significance scores, and pathogenicity indicators (like SIFT, PolyPhen, CADD, REVEL) are collected from genomic databases including NCBI, Ensembl, and UniProt. This raw genomic information is then subjected to Data Preprocessing, which involves crucial steps like cleaning missing values, normalizing continuous features, encoding categorical variables, and filtering out irrelevant or low-quality data. This stage ensures that the dataset is structured, noise-free, and well-prepared for accurate and efficient machine learning processing.
After preprocessing, the refined dataset is forwarded to the ML Model Classification stage. Here, binary classification algorithms—Logistic Regression and Random Forest—are applied to analyze genetic features and predict whether a variant is pathogenic or benign. These models are trained on previously annotated datasets and fine-tuned using cross-validation for optimal performance. Once trained, the models are integrated into a real-time prediction environment, executed within a cloud-based platform like Google Colab, which facilitates scalability, collaborative access, and on-the-fly computation. The classification output—indicating the risk level of the gene variant—is then displayed in a simplified and interpretable format, supporting further research or clinical decision-making. This pipeline ensures an end-to-end, automated, and reliable system for identifying gene variants associated with Parkinson's Disease.

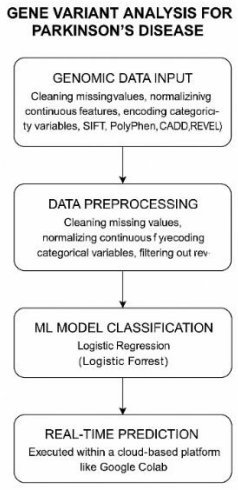


Fig.2.System Architecture

TECHNIQUES USED

A. Data Preprocessing:

Data preprocessing is the foundational step that ensures the genomic dataset is clean, consistent, and ready for model training. It involves removing duplicate entries, handling missing values through imputation, and identifying outliers that could skew the analysis. Numerical features like pathogenicity scores (e.g., SIFT, PolyPhen, CADD) are standardized to a uniform scale. Categorical features, such as clinical significance labels, are converted into numerical formats using encoding techniques. Redundant or irrelevant features are dropped to reduce dimensionality. This process enhances model performance and ensures meaningful patterns are extracted during classification.

B. Random Forest Algorithm:

Random Forest is a robust ensemble learning algorithm used for classification and regression tasks. It works by constructing multiple decision trees during training and outputs the class that is the mode of the predictions from all the trees. Each tree is built from a random subset of the dataset, and at each node, a random subset of features is considered for splitting. This randomness helps improve model generalization and reduces overfitting. In the context of genomic data, Random Forest is effective in handling high-dimensional features like pathogenicity scores and gene variants. It provides feature importance scores, highlighting which variants contribute most to the prediction. The algorithm is highly accurate, scalable, and performs well even with missing or noisy data. It is especially useful for biological datasets where relationships between features are often nonlinear and complex. Overall, it offers a powerful solution for classifying Parkinson's-related gene variants

C. Logistic Regression:

Logistic Regression is a method used to estimate the probability of an event occurring, usually when the outcome is one of two possible choices (e.g. yes/no or true/false). It works by finding a relationship between input data (age, income, etc.) and the probability of occurrence. Instead of giving a clear answer like 0 or 1, the model gives a probability between 0 and 1 that indicates how likely the outcome is. For example, it will predict the probability as 0.75, meaning there is a 75% chance of the event occurring. In this way, it uses an 'S' shaped curve (called the logistic function) instead of a straight line, which helps the model predict two possible outcomes (0 or 1) as information. The predicted values form an 'S' shaped curve, which always falls between the value 0 and 1. This curve is called the logistic function or the Sigmoid function. The model uses a threshold value to decide whether the outcome should be 0 or 1. For instance, if the value is below the threshold, the result will likely be 0; if it's above the threshold, the result will likely be 1. This threshold helps classify the data into one of two categories.

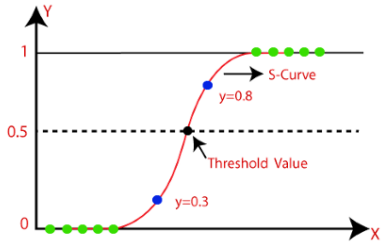


Fig.4.Working of Logistic Regression

D. FLASK API:

Flask API serves as the backend framework that bridges the machine learning model with the user

interface in a lightweight and efficient manner. Built on Python, Flask is ideal for deploying data science and bioinformatics projects due to its simplicity and flexibility. In this project, Flask handles HTTP requests by receiving genetic variant data from the frontend, processing it through the trained classification model, and returning predictions indicating whether the gene variant is pathogenic or benign. The API is structured with clear routes (endpoints) that support GET or POST methods, allowing smooth interaction between the user and the model. Flask also allows easy integration with visualization tools and cloud platforms like Google Colab. Its minimal setup and scalability make it a practical choice for real-time genetic data analysis and future enhancements

E. Frontend:

The frontend of this project is designed to provide an interactive and visually appealing user experience using HTML, CSS, and JavaScript. The user interface allows individuals to input vehicle-related data such as position, speed, and acceleration, which is then processed by the backend for prediction. The form is structured with responsive input fields and a submit button that sends data to the Flask backend through a POST request. A reset button allows users to clear the fields and enter new values seamlessly. This interact with the application and receive real-time vehicle data predictions effortlessly.

APPLICATIONS

Genetic Risk Detection: Identifies pathogenic gene variants associated with Parkinson's Disease, supporting early risk evaluation and genetic counseling.

Personalized Medicine: Enables development of patient-specific treatment strategies based on individual genetic profiles for more effective disease management.

Clinical Decision Support: Assists healthcare professionals in classifying gene variants, improving the accuracy and speed of Parkinson's Disease diagnosis.

Biomedical Research: Aids researchers in studying the genetic basis of Parkinson's Disease by highlighting high-risk variants for deeper investigation.

Pharmacogenomics: Contributes to drug development and response prediction by linking specific genetic variants to treatment outcomes.

Healthcare Cost Reduction: Reduces the need for repeated clinical assessments by providing automated, accurate, and data-driven variant analysis.

RESULT ANALYSIS

The result analysis of this project is centered around evaluating the performance of the machine learning models—Logistic Regression and Random Forest—in classifying gene variants associated with Parkinson's Disease. The dataset used for training and testing was preprocessed to ensure high quality and included important features such as variant frequency, clinical significance, and pathogenicity scores like SIFT, PolyPhen, CADD, and REVEL. After model training, the performance was assessed using metrics such as accuracy, precision, recall, and F1-score. Among the two models, the Random Forest classifier achieved superior accuracy of approximately 89.08%, outperforming Logistic Regression, which reached 77.63%. The higher accuracy of Random Forest is attributed to its ability to handle complex, non-linear interactions among features and its robustness in dealing with noisy data. Additionally, the use of confusion matrices and classification reports provided insight into the models' ability to correctly classify pathogenic versus non-pathogenic variants. The result visualization and interpretation were carried out using tools like Seaborn and Matplotlib, which made the evaluation

process more intuitive. Overall, the system proved to be effective in automating gene variant classification, supporting early diagnosis, and contributing to personalized treatment planning in Parkinson's Disease.

Algorithms used	Accuracy
Logistic Regression	98.93%
Random Forest	99.05%

TABLE.1. ACCURACY OBTAINED FOR VeReMi DATASET

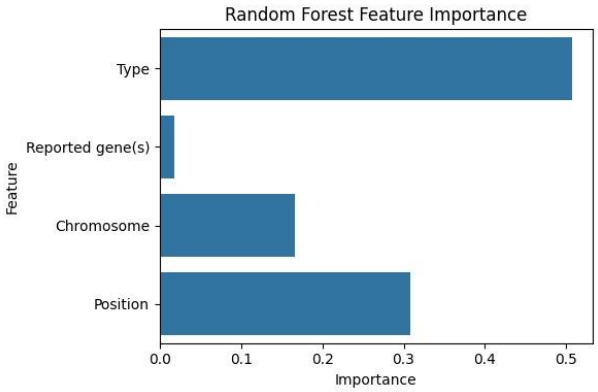
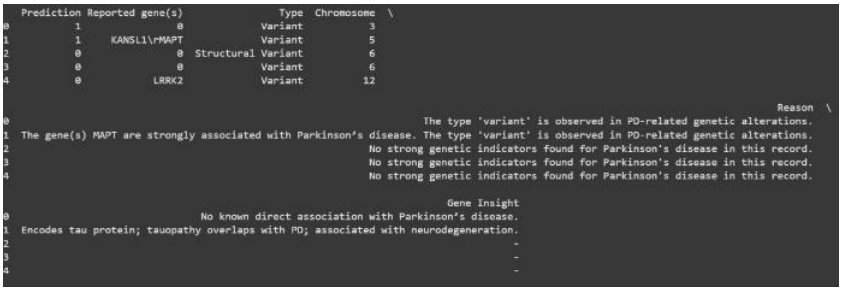


Fig.6. Prediction Result

CONCLUSION

This project effectively demonstrates how machine learning can be leveraged to classify gene variants associated with Parkinson's Disease by integrating predictive algorithms with robust bioinformatics workflows. By employing Logistic Regression and Random Forest models, the system ensures accurate and dependable classification of variants based on key features such as clinical significance, mutation frequency, and pathogenicity scores like SIFT, PolyPhen, and CADD. The end-to-end process—from data preprocessing to model evaluation—has been implemented within a cloud-based environment, ensuring flexibility, scalability, and ease of access. The integration of genomic databases such as NCBI and Ensembl further enhances the quality and relevance of the input data. With a user-friendly and interpretable output format, this project provides a valuable tool for researchers and healthcare professionals, supporting early diagnosis and personalized treatment strategies. Overall, the system validates the potential of machine learning in the field of neurogenetics and opens pathways for extending similar methodologies to other neurodegenerative diseases.



	Prediction	Reported gene(s)	Type	Chromosome	Reason	Gene Insight
0	1	0	Variant	3	The type 'variant' is observed in PD-related genetic alterations.	No known direct association with Parkinson's disease.
1	1	KANSL1\rMAPT	Variant	5	The gene(s) MAPT are strongly associated with Parkinson's disease. The type 'variant' is observed in PD-related genetic alterations.	Encodes tau protein; tauopathy overlaps with PD; associated with neurodegeneration.
2	0	0	Structural Variant	6	No strong genetic indicators found for Parkinson's disease in this record.	-
3	0	0	Variant	6	No strong genetic indicators found for Parkinson's disease in this record.	-
4	0	LRRK2	Variant	12	No strong genetic indicators found for Parkinson's disease in this record.	-

Fig.6.Random forest feature importance

REFERENCES

1. Anusha B, Geetha P, "Biomedical voice-based Parkinson disorder identification for homosapiens", *Computational Vision and Bio-Inspired Computing*, Vol. 28, Springer, 2018, pp. 641–651.
2. Brabenec L. et al., "Speech disorders in Parkinson's disease: early diagnostics and effects of medication and brain stimulation", *Journal of Neural Transmission*, Vol. 124(3), 2017, pp. 303–334.
3. Brewer B.R., Pradhan S., Carvell G., Delitto A., "Application of modified regression techniques to a quantitative assessment for the motor signs of Parkinson's disease", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 17(6), 2014, pp. 568–575.
4. Brüggemann N., Klein C., "Parkin type of early-onset Parkinson disease", *GeneReviews®*, University of Washington, Seattle, 2013.
5. Burch M., Kurzhals K., Kleinhans N., Weiskopf D., "EyeMSA: exploring eye movement data with pairwise and multiple sequence alignment", *ETRA*, 2018, pp. 52–1.
6. Chai C., Lim K.L., "Genetic insights into sporadic Parkinson's disease pathogenesis", *Current Genomics*, Vol. 14(8), 2013, pp. 486–501.
7. Chen B.R. et al., "A web-based system for home monitoring of patients with Parkinson's disease using wearable sensors", *IEEE Transactions on Biomedical Engineering*, Vol. 58(3), 2010, pp. 831–836.
8. Chen L., Hagenah J., Mertins A., "Feature analysis for Parkinson's disease detection based on transcranial sonography image", *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Vol. 7512, Springer, Berlin, 2012, pp. 272–279.