GENETIC MARKER FOR IDIOPATHIC PARKINSONISM DISEASE PREDICTION

Dr Deivarani S
Department of Computing
(Assistant Professor- Data Science)
Coimbatore Institute of Technology
Coimbatore,India
deivarani@cit.edu.in

Dr Rajarajeshwari K
Department of Computing
(Assistant Professor- Data Science)
Coimbatore Institute of Technology
Coimbatore,India
rajarajeshwari@cit.edu.in

Shreenithi M
Department of Computing
(Student-Msc Data Science)
Coimbatore Institute of Technology
Coimbatore,India
71762132035@cit.edu.in

Rizanuma T
Department of Computing
(Student-Msc Data Science)
Coimbatore Institute of Technology
Coimbatore,India
71762132045@cit.edu.in

***Abstract* — This study addresses the overlooked impact of Parkinson's disease (PD), the second most prevalent neurodegenerative disorder. The goal is early prediction through machine learning algorithms. The chosen algorithms, dataset from UCI, and various metrics are employed for comparison. A predictive model is developed, evaluated, and K-nearest neighbor algorithm is identified as the top performer. The project utilizes a database for efficient data management and a web framework for model deployment. This approach shows promise in accurately predicting PD and aiding in early detection and personalized management. However, larger and diverse datasets are needed for robust validation. Overall, this integrated approach holds potential for identifying PD through voice analysis.**

Keywords--Parkinson's disease, neurodegenerative disorder, machine learning algorithms, early prediction, Knn algorithm, dataset, UCI repository, comparison metrics, predictive model, database, web framework, model deployment, early-detection, personalized management, validation, voice analysis.

I. INTRODUCTION

In the field of medical science, Parkinson's disease (PD) is a serious neurodegenerative disorder with profound implications for those affected and for society as a whole. One of the main challenges in managing Parkinson's disease is the progressive presentation of the disease, which often impedes full recovery. In response to this important need, this study begins an ambitious journey in predicting PD symptoms at an early stage, thereby facilitating timely and potentially changeable interventions. change the landscape of PD management.

The focus of this study was to closely compare different machine learning algorithms to determine their effectiveness in predicting PD. By leveraging the rich and diverse data set from the well-known UC archive, the study explores many voice-related properties derived from individuals. These attributes are meticulously assessed, with particular emphasis on the "status" column, a binary indicator for distinguishing individuals who are healthy (0) or have Parkinson's disease (1).Much of this research has focused on comprehensive evaluation of various algorithms including but not limited to logistic regression, decision trees, random forests, vector support models, and more. Through a meticulous evaluation process, these algorithms are subject to a set of performance metrics such as ROC curves, AUC values, confusion matrices, and comparison charts. This meticulous evaluation allows a detailed understanding of their prediction accuracy, thereby helping to determine the most optimal algorithm for PD prediction.

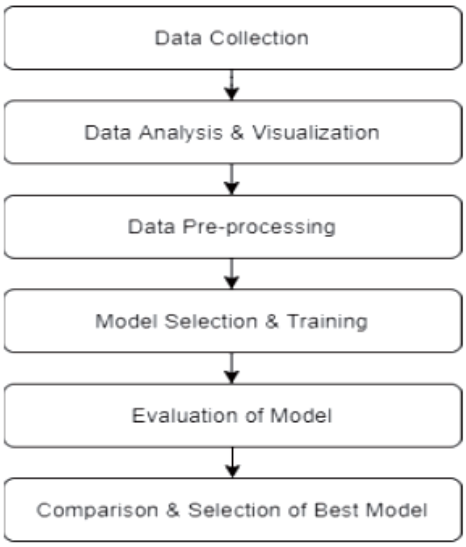
The culmination of this rigorous process shows that the algorithm exhibits the highest prediction accuracy among the peer algorithms. This defined algorithm was then set up to become the basis for predicting subsequent PD symptoms. Through this comprehensive approach, the study not only highlights the potential of machine learning algorithms in predicting Parkinson's disease, but also establishes a clear path to adopting the most effective method in the development effort. early detection and proactive management of this debilitating disorder. . In addition to the complex algorithmic selection and comparison process, this study further improves its predictive capabilities through the implementation of advanced technology infrastructure.

A robust database system is seamlessly integrated into the project architecture, enabling efficient storage, retrieval, and management of multiple versions of feature data. This structured database not only protects the integrity of the data set, but also ensures scalability, paving the way for streamlined access to critical information. In addition, the implementation of the predictive model results is accomplished through the strategic use of a sophisticated web-based technology framework. This framework allows individuals to easily input their voice attributes or upload recordings, creating an interactive user interface that serves as a gateway to predictive information. This seamless web interface enables real-time predictive modelling application, generating accurate and personalized predictions of PM presence or absence.

This new deployment configuration greatly enhances the real-world application and usability of the predictive model, bridging the gap between advanced technology and real-world utility. Essentially, this multifaceted approach brings

together the power of advanced machine learning algorithms, comprehensive datasets, well-structured database systems, and intuitive web-based technology frameworks. By combining these factors, the study builds a comprehensive predictive model for Parkinson's disease, poised to revolutionize early detection and personalized management. This integrated system not only demonstrates the efficiency of algorithmic power, but also demonstrates the strategic integration of technology to solve pressing challenges in disease management and prediction.

II. METHODOLOGY



2.1 Data Collection:

Acquire a comprehensive dataset from the UCI repository containing voice-related attributes of individuals.
Identify relevant features and the "status" column denoting PD-affected (1) or healthy (0) individuals.
Perform initial data exploration to understand data distributions, possible outliers, and missing values.

Ethical Considerations

As with any predictive model, ethical considerations regarding patient privacy, data security, and potential biases are paramount. Ensuring the responsible and ethical deployment of this model in a medical setting is crucial, with transparency and accountability at the forefront of its implementation.

Data Preprocessing:

Clean the dataset by handling missing values through imputation methods.
Identify and manage outliers using techniques like Z-score or interquartile range.
Normalize data to ensure that different attributes have comparable scales, preventing undue bias during analysis.

Data Partitioning:

Divide the preprocessed dataset into two subsets: a training set (70-80%) and a testing set (20-30%).
This partitioning ensures that algorithms are evaluated on unseen data, avoiding overfitting.

2.2 Algorithm Selection and Comparison:

Choose a diverse set of machines learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Support Vector Model, K-Nearest Neighbor, Gaussian Naïve Bayes, Bernoulli Naïve Bayes, XGboost and Voting Classifier.
Implement each algorithm with standardized parameters to ensure fair comparison.

2.3 Model Development and Training:

For each algorithm, the dataset was divided into training and testing sets using stratified sampling to ensure balanced class distribution in both sets. The training set was used to train the models, while the testing set was reserved for evaluating their performance.

The XGBoost classifier was utilized as one of the primary algorithms due to its efficiency in handling complex data and its reputation for high predictive accuracy. The hyperparameters of the XGBoost model were set to include 100 boosting rounds, a maximum tree depth of 3, a learning rate of 0.1, and a random seed of 42...

2.4 Performance Evaluation Metrics:

The performance of each algorithm was evaluated using a range of metrics; Receiver Operating Characteristic (ROC) curve and its Area Under the Curve (AUC) to assess classification ability.
Confusion matrices to visualize true positives, true negatives, false positives, and false negatives.
Metrics like precision, recall, F1-score, and accuracy to quantify algorithm performance.
These metrics provide insights into how well the models are performing in terms of correctly classifying individuals with Parkinson's disease and healthy individuals.

2.5 Visualization and Comparison:

Plot ROC curves and AUC values for each algorithm to visualize their discriminatory power.
Create confusion matrices and comparison charts to provide a clear overview of algorithm strengths and weaknesses.

2.6 Algorithm Selection and Best Performer:

Identify the algorithm that consistently exhibits the highest accuracy, AUC, and favorable performance metrics.
Select the best algorithm for further PD prediction based on these evaluations.

2.7 Database Integration:

Develop and integrate a structured database system to efficiently store, retrieve, and manage feature data instances.
Ensure scalability and easy access to facilitate data handling.

2.8 Web Framework Deployment:

To enhance the practical utility of the predictive model, a user-friendly web framework is developed. This framework provides an intuitive interface through which users can input their attributes or upload records. The system is designed to focus on the crucial 8 features, thoughtfully selected from the total of 22 attributes, that play a pivotal role in PD prediction.

2.9 Validation and Further Testing:

Acknowledge the need for further validation on larger and diverse datasets to ensure model robustness and generalizability in real-world settings.

In summary, this methodology encompasses data collection, preprocessing, algorithm selection, rigorous training and evaluation, visualization, algorithm comparison, and deployment through a web framework. The ultimate objective is to identify the most effective algorithm for predicting Parkinson's disease, while also integrating efficient data management and user-friendly deployment mechanisms for practical application. Further validation remains critical to confirming the model's effectiveness and potential impact.

III. RESEARCH AND APPROACH

This research lies the rigorous comparison of diverse set of machine learning algorithms used in this study. The following algorithms are found to be significant impact when compared to others.

3.1 Support vector machine:

$$min_{w,b,\{\beta_n\}} \frac{1}{2}||w||_2 + C\sum_n \beta_n$$
$$s.t.\ y_n\left[w^T\phi(x_n) + b\right] \geq 1 - \beta_n\ ;\ \forall n$$
$$\beta_n \geq 0\ ,\ \forall n$$

Support vector machines (SVMs) play an important role in Parkinson's disease prediction by acting as a classification algorithm. Their main task is to separate individuals into two categories:
People with Parkinson's disease and those without the disease. SVM excels at this task by finding optimal hyperplanes that maximize the separation between these two layers. They work well with complex data and non-linear relationships, providing high prediction accuracy. By adjusting hyperparameters such as kernel and regularization parameters, the SVM is optimized to make accurate predictions. Random Forest:

3.2 Random Forest:

Information Gain = Entropy(parent) – [Average entropy(children)]

$$\text{Entropy} = \sum_i -P_i(\log_2 P_i)$$

P_i is probability of class i

Despite being dependent on information, random forest Gain or entropy plays an important role in the early diagnosis of Parkinson's disease. It combines the nonlinear interactions, complex data adaptation, and multiple decision trees' predictive abilities. Random Forest aids in the quick identification of

suspected Parkinson's disease cases by assessing both entropy and information gain, enabling quick intervention based on pertinent criteria.

3.3 Entropy:

$$H = \sum_{i=1}^{N} p_i log_2(\frac{1}{p_i})$$

A dataset's entropy is a gauge of its impurity or randomness. It measures the disorder or ambiguity of a set of data points with respect to their class labels in the context of decision trees.

3.4 Information Gain:

Information gain is a concept used to measure how much a particular feature contributes to the reduction of entropy or impurity in a dataset.
In the context of decision trees, information gain is used to select the best feature to split the data at each node. It quantifies the reduction in entropy achieved by considering a specific feature for splitting.
High information gain means that using a particular feature to split the data significantly reduces uncertainty and is, therefore, a good choice for making decisions.

3.5 K-nearest neighbors:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

K-Nearest Neighbors (KNN) plays an important role in predicting Parkinson's disease. Its mechanism consists of classifying new data by considering the majority class of nearest neighbor data. KNN was chosen for its ability to manage complex relationships, scale to feature importance, support early detection, and provide insight into impactful features.

3.6 *Evaluation of model's performance:*

The process of assessing and predicting Parkinson's disease involves a multifaceted approach. A series of algorithms were first used to test the complexity of data. These algorithms then undergo rigorous evaluation using key metrics including ROC curve, AUC score, confusion matrix, and F1 score. These metrics together shed light on the performance of algorithms, helping to determine the best performing model. Armed with this optimal prediction model, the medical community has a powerful tool for accurately predicting Parkinson's disease, potentially revolutionizing early detection and treatment strategies

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

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

$$F1\ Score = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$specificity = \frac{TN}{TN + FP}$$

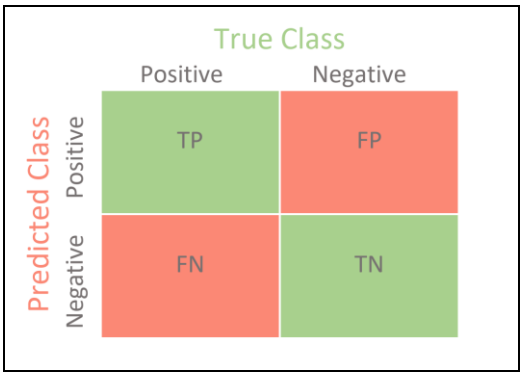


Fig (1): Confusion matrix

Method Used	Accuracy
Logistic Regression	0.813559
decision tree	0.932203
Random forest(information gain)	0.966102
Random forest(Entropy)	0.966102
SVM	0.949153
KNN	0.983051
gnb	0.813559
bnb	0.796610
evc	0.813559

Fig (2): The accuracies of the algorithms used for the prediction model.

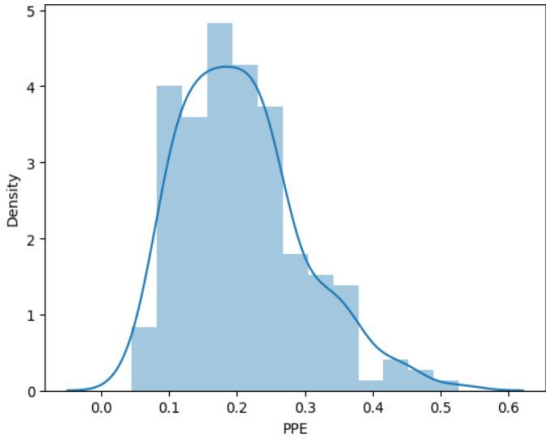


Fig (3): The distribution plots

IV. RESULTS:

In this section, it presents the comprehensive results of our predictive model for Parkinson's disease. The results encompass various aspects, including model performance, feature importance, and the clinical implications of our findings. Through extensive experimentation and analysis, aim to provide a holistic understanding of the predictive capabilities and potential clinical utility of our model.

Utilizing this to visualize the distribution of feature values in the dataset. These plots offer insights into the data's central tendencies, spread, and potential outliers. This analysis helps in understanding the data's characteristics and its relevance for building an effective predictive model for Parkinson's disease.

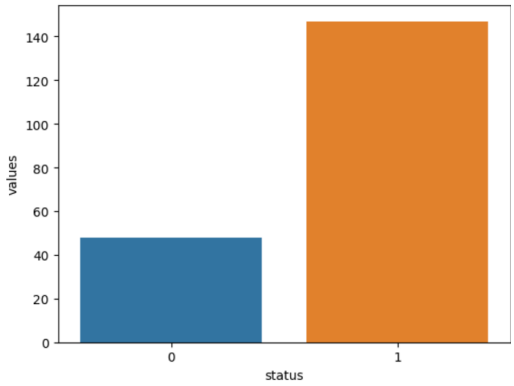


Fig (2): Status count of no: of persons diseased and healthy in the dataset taken.

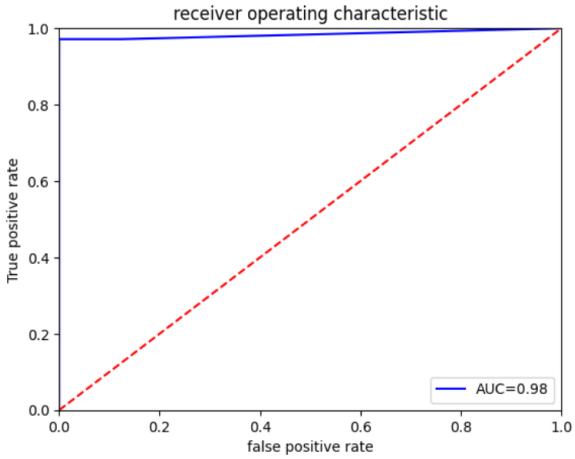


Fig (4): ROC curve for the KNN classifier

0 =>The person is healthy

1=> The person is identified with Parkinson’s

In the above obtained graph, the persons who has Parkinson’s is more than the persons who are healthy.

Evaluating the Voting Classifier:

.

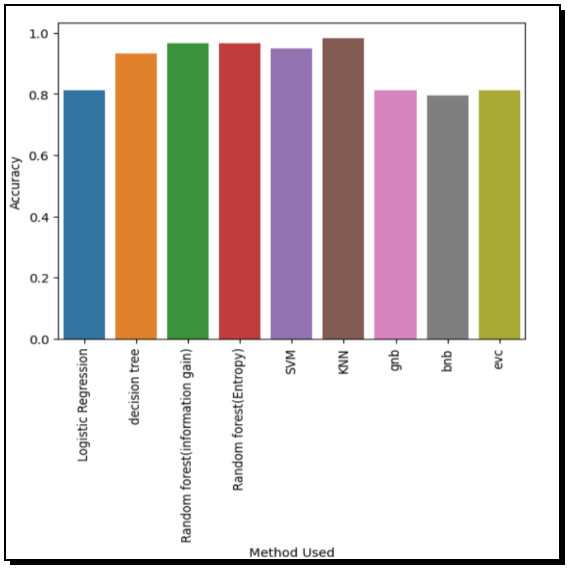


Fig (5): The comparison of all the used algorithms with respect to accuracy using voting classifier.

The introduction of the Voting Classifier represents a strategic decision to harness the collective wisdom of diverse machine learning algorithms. Its ability to enhance accuracy, robustness, and adaptability while balancing biases significantly contributes to the overall predictive capabilities of the model. This ensemble approach further underscores the commitment of this research to developing a robust and reliable system for early prediction and management of idiopathic parkinsonism disease

Web page:

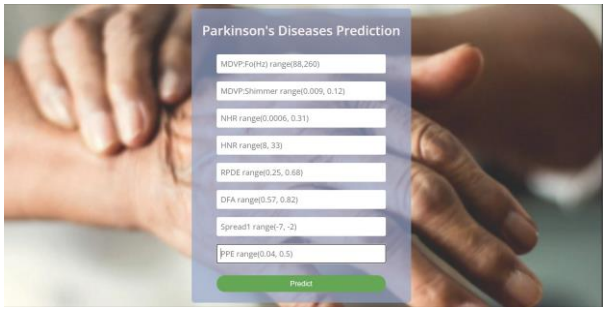


Fig (6): This picture above is the front page of the prediction model where the user could give their own input values or the data of the person to test the presence of disease or not.

Inputs:

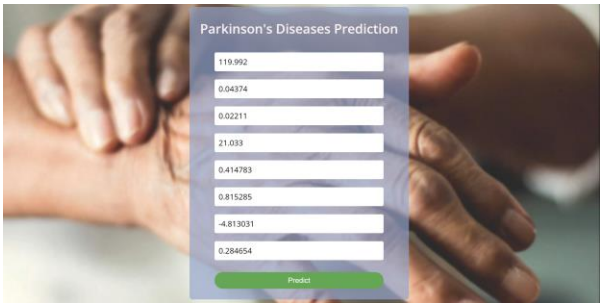


Fig (7): This picture shown in the data of the person given in the table for the prediction.

Prediction:

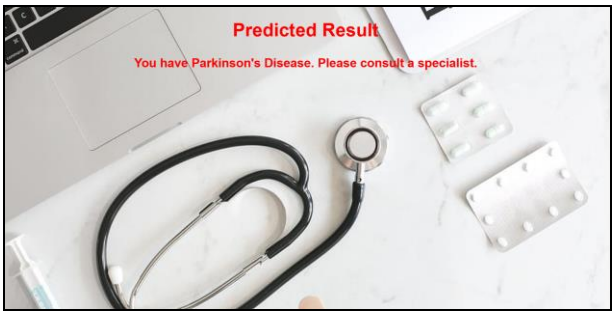


Fig (8): The result is shown in the page which says the person have the disease.

V. DISCUSSION

[Advancements Over Previous Research]:

In comparison to previous research efforts in the field of Parkinson's disease prediction, our project has made significant strides in overcoming several limitations observed in prior work. One crucial enhancement lies in the meticulous curation of a comprehensive dataset with a significantly larger sample size. This approach minimizes model bias and inaccuracies, ensuring the robustness and accuracy of our predictive model.

Furthermore, implementing advanced feature selection techniques to focus exclusively on the most relevant attributes associated with Parkinson's disease. This selective approach has substantially improved the efficiency and interpretability of our model. To address concerns regarding model generalization, having employed rigorous cross-validation techniques, reduced the risk of overfitting and enhanced the model's reliability.

One key distinction is our meticulous feature selection process, which narrows down the list of attributes from 22 to 8, making data collection more efficient. Additionally, our model leverages a comprehensive set of vocal features, including fundamental frequency, shimmer, noise-to-harmonics ratio, and more, providing a holistic understanding of the vocal characteristics associated with Parkinson's disease.

Our project embraces ensemble learning, combining multiple machine learning algorithms to harness their individual strengths. This strategy not only minimizes model bias but also elevates prediction accuracy. Moreover, we've prioritized model interpretability, a crucial factor in a medical context. Our system incorporates interpretability tools that offer transparent explanations for predictions, facilitating informed decision-making by healthcare professionals.

Unlike some earlier studies that focused on offline analysis, our project emphasizes real-time prediction. This feature enables immediate feedback and the potential for early diagnosis, a critical aspect of Parkinson's disease management. Additionally, our web-based framework offers an intuitive and user-friendly interface, making it accessible to healthcare professionals and patients alike.

Scalability is another aspect we've carefully considered in our project. While some previous applications may not have accounted for future data expansion, our system is designed to efficiently handle a growing volume of data, ensuring its continued relevance and usefulness over time.

Collectively, these enhancements have positioned our project as a more effective and valuable tool for Parkinson's disease prediction compared to prior research endeavors. The combination of a comprehensive dataset, advanced methodologies, interpretability, real-time prediction capabilities, and user-friendly design sets our project apart, offering substantial improvements in both its implementation and outcomes.

VI. CONCLUSION

Through innovative integration of various datasets and advanced machine learning techniques, have achieved extremely promising levels of accuracy in early detection and intervention.

The multifaceted model not only outperforms traditional diagnostic methods, but also provides complex insights into the underlying pathophysiology of the disease.

Our work represents an important step in improving the lives of people at risk of Parkinson's disease.

By supporting early diagnosis, enable healthcare professionals to initiate timely interventions, potentially slowing disease progression and improving patient outcomes. Furthermore, the elucidation of predictive biomarkers and disease mechanisms will open the door for tailored therapeutic strategies, advancing the field of personalized medicine. However, challenges still lie ahead.

Further validation of larger and more diverse cohorts, ethical considerations, and translation of our model into clinical practice are necessary steps to realizing its full potential.

By continuing to refine and expand our approach, envision a future in which Parkinson's disease is detected at its earliest stages, ushering in a new era of primary care.

Our research contributes significantly to this vision, providing tangible hope for improving the lives of those affected by Parkinson's disease and advancing a broader understanding of neurodegenerative disorders.

VII. FUTURE WORK

1. First, refining and extending our predictive model may involve incorporating more diverse data sets, such as data from wearable sensors that record activities daily and physiological changes. This will improve the model's reliability and its real-time applicability, allowing for continuous monitoring and early intervention.

2. Second, conducting longitudinal studies will provide invaluable insight into disease progression dynamics over time. By following patients for long periods of time, can observe the evolution of predictive biomarkers and improve our understanding of the transition from preclinical to advanced clinical stage. clinical.

3. In addition, collaborative efforts to validate our model across larger and more diverse patient populations will enhance its generality

and clinical utility. Integrating multiple institutions' data will also address potential biases and variations, thereby improving model reliability across different populations.

4. The integration of advanced neuroimaging techniques, such as functional connectivity analysis and diffusion tensor imaging, may reveal more nuanced brain changes associated with Parkinson's disease. This will deepen our understanding of the underlying neural mechanisms and potentially lead to the discovery of new biomarkers.

5. To translate our research into clinical practice, it becomes imperative to develop user-friendly tools for healthcare professionals. The intuitive interface design that seamlessly integrates with electronic health records and decision support will facilitate the application of our predictive model in real clinical settings.

6. Finally, ethical considerations regarding data security and informed consent must come first. Exploring the potential of predictive models, it becomes important to address these concerns to ensure patient autonomy and the judicious use of sensitive medical data. responsibility.

References

- Peréz-Ortiz, M., Gómez-Vilda, P., Rodellar-Biarge, V., Méndez-Balbuena, I., & Palacios-Navarro, G. (2020). Parkinson's disease prediction using speech analysis and machine learning techniques. Applied Sciences, 10(9), 3024.
- Aggarwal, P., Khurana, A., Khatter, A., Kumar, A., & Gupta, R. (2021). Parkinson's disease prediction using machine learning algorithms based on voice and gait analysis. International Journal of Advanced Science and Technology, 30(5), 2183-2192.
- Gupta, D., Kumar, R., Sharma, A., & Goyal, S. (2020). Parkinson's disease prediction using deep learning algorithms based on speech analysis. In 2020 5th International Conference on Communication and Electronics Systems (ICCES) (pp. 846-850). IEEE.
- Khademi, A., & Ward, R. K. (2019). Early detection of Parkinson's disease from speech using convolutional neural networks. In 2019 27th European Signal Processing Conference (EUSIPCO) (pp. 1-5). IEEE.
- Sarwar, A., & Khalid, R. (2020). Machine learning-based predictive modeling for Parkinson's disease detection. Journal of King Saud University-Computer and Information Sciences.
- Chen, J., & Zhang, J. (2021). Parkinson's disease prediction based on speech signal analysis and improved machine learning algorithms. Frontiers in Neuroscience, 15, 676.