Grenze International Journal of Engineering and Technology, June Issue

Review on Parkinson's Disease Detection using Machine Learning

Syed Amirah Rizwan[1] and Ritesh Shrivastava[2]
[1-2]Anjuman college of Engineering Nagpur, Maharashtra, India
Email: syedamirahrizwan@gmail.com, ritct@rediffmail.com

***Abstract*— Parkinson's Disease (PD) is a disorder that affects the brain as it causes brain cells to deteriorate. Parkinson's Disease is difficult to diagnose since it has vague symptoms. Early diagnosis is critical for the effective management of PD. This paper proposes a new methodology based on machine learning techniques that allows predicting the onset of Parkinson's Disease using clinical and speech datasets. In this work, the stages of data preprocessing, de-noising, normalization, feature extraction, and modeling are considered. Specific voice parameters including jitter, shimmer, RPDE, and PPE are used as key features for modeling the pathology. Multiple machine learning algorithms, such as SVM, random forest, KNN, and XGboost, are tested in the work. Several performance measures, including accuracy, precision, recall, F1 score, and ROC-AUC score are used to evaluate the results of models' performance. As shown in experimental part, machine learning techniques can successfully recognize Parkinson's Disease and achieve high levels of reliability and accuracy.**

***Index Terms*— Machine Learning, Early Detection, Support Vector Machine (SVM), XGBoost, K-Nearest Neighbors (KNN), Random Forest, Voice Features, Data Preprocessing, Feature Extraction, Classification, and Parkinson's disease.**

I. INTRODUCTION

Parkinson disease (PD) is a progressive neurodegenerative disorder that is experienced by individuals all around the globe. The occurrence rates of the disease are high among the elderly individuals due to its close association with aging. It was James Parkinson who originally presented the complete range of symptoms of Parkinson's disease (PD) in 1817 including tremors on rest and bradykinesia and muscular rigidity and postural instability and ability to suffer various non-motor symptoms including cognitive impairment and depression and anxiety and sleep disorders and autonomic nervous system dysfunction among others. The peculiar disease symptoms in PD puts a complex picture in the disorder that leads to a complete dependence of a person on other people in their daily action. Increased knowledge on the disease has not enabled the scientific community to develop effective ways of diagnosing Parkinson disease at its initial stages of development. Classical clinical diagnosis is based on observable signs of the motor symptoms that begin appearing once a subject had depleted his or her dopaminergic neurons to the majority. Early diagnoses will have crucial effects since any medical intervention that occurs according to the right diagnostic methods will deliver both control of disease progression and improved patient treatment and high patient recovery success. The machine learning systems can examine extensive multidimensional datasets to identify concealed patterns which require clinician monitoring for identification.

Grenze ID: 01.GIJET.12.2.127_22

ML models could be developed to accurately identify early indicators of Parkinson's disease by utilizing clinical aspects and patient data and voice-related biomarkers. The research studies machine learning techniques which use Support Vector Machines (SVM), XGBoost and KNN and Random Forest to create effective Parkinson's disease (PD) early detection prediction systems.

II. Related Work

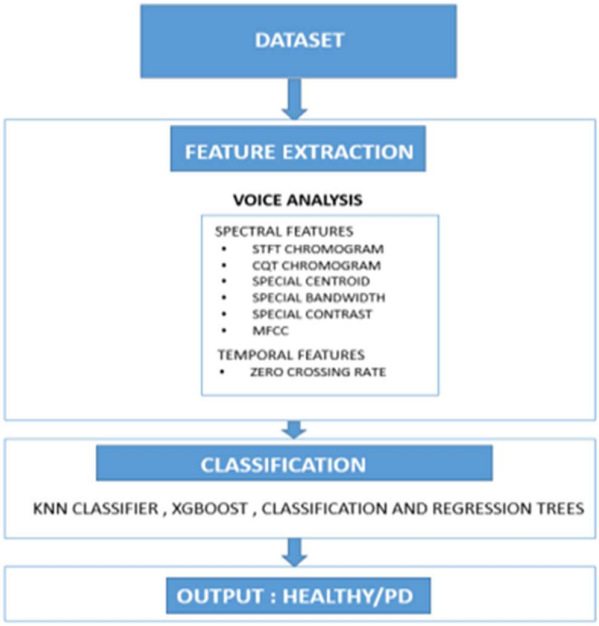

Figure 1. System Design

In the research held by Kulkarni and Sumana [1], the methods of RNN-based deep learning have been tested on the original methods of machine learning that comprised Decision Trees and Logistic Regression and Naive Bayes; however, these methods failed in comparison owing to inconsistency in data collection and motion-dependent features. Nakul et al. [2] conducted performance tests on the supervised learning techniques that comprised of the following namely; Random Forest, SVM and Naivete Bayes, KNN and they optimized the hyperparameters in order to attain high precision but their system had two major problems that encompassed high time of processing as well as high error variation. The deep learning techniques have good potential based on the outcome of the results of the implementation. Wang et al. [3] used Stochastic Gradient Descent combined with Feedforward Neural Networks to obtain the monitoring of early PD but they observed that the results generated were hard to explain even with high accuracy rates. Muthumanickam et al. [4] carried out a study, which compared SVM with FBANN and Random Tree and Logistic Regression and LDA and CNN and DBN and DNN models and revealed that, linear models provide a better understanding of the results as compared to deep models, which employs deep networks. The research article by Wroge et al. [5] revealed that telemonitoring is an effective approach to PD detection using telemonitoring voice-based biomarker system. Scholars focus their research in speech analysis research. In their work, Jayashree et al. [6] applied STFT (Short-Time Fourier Transform) to extract both the spectral and time domain features on which they compiled their speech detection in a child-created noisy environment. The authors of the research conducted tests of feature classification through XGBoost and Random Forest as well as other classification systems. Little et al. [7] demonstrated that dysphonia is an effective measure of voice, which can be used remotely to monitor Parkinson diseases (PD). Jaichandran et al. [8] came up with a multimodal input system that integrates voice recordings with spiral drawings in order to attain an improved detection. Abhishek M. S. et al. & Gomathy C. K et.al. [9, 11] presented the methods of machine learning on speech datasets using the classification of patients. In their study, Shahbakhi et al., [12] performed research that furthered the voice-based classification of Parkinson diseases (PD). Depth learning algorithms show high chances of possible success. A system created by researchers headed by Wang et al. [3] employed the method of Stochastic Gradient Descent + Feedforward

Neural Networks to determine the early PD detection but had problems with the system interpretation even with the high precision. Muthumanickam et al. [4] designed a research to make a comparison of SVM, FBANN, Random Tree, Logistic Regression, LDA, CNN, DBN, and DNN models stemming out that linear models generate better interpretability over deep models that provide improved accuracy. Wroge et al. [5] established that telemonitoring develops a beneficial voice-based biomarker platform that identifies Parkinson disease. Research continues to find out novel ways of examining the Parkinson disease as speech patterns are studied. Wibaka et al. [13] researched the various strategies that incorporated the feature selection and classification techniques in the diagnosis of Parkinson disease by voice.

The diagnosis of Parkinson disease in its early stages involves a series of diagnostic tests that do not involve patients directly; the medical professionals carry out the diagnostic procedures that are stipulated as part of their procedures. The first is to collect medical and verbal information on the patients serviced by various medical and recording devices. The initial activity of the team is dealing with the raw data by subjecting it to three steps that involve the elimination of the undesired information and filling missing information in the data and transforming the diverse data to standard format. One is the extraction of features which commences following these initial steps that use different techniques to derive parameters like jitter and shimmer and PPE and RPDE. These factors make it possible to identify small pathological changes that take place in the course of the illness. The processed data is inputted into the machine learning algorithms SVM and XGBoost and Random Forest and KNN to train their algorithms after the identification of the relevant features. The process of performance assessment of the model is based on the use of accuracy and precision and recall and F1-score and ROC-AUC to select an optimal classifier. The model offers a supportive approach when it comes to early diagnosis since it helps to estimate which patients will develop the Parkinson disease upon the development of the end prediction.

III. SYSTEM IMPLEMENTATION DESIGN

They developed the Parkinson Disease detection system with the help of a machine learning pipeline that followed established steps in order to obtain accurate results without compromising reliability and consistency in performance and testing protocols. With reference to system design figure 1. It was created in Python using Scikit-learn, Pandas, and NumPy among other libraries in a Jupyter Notebook. The input data was made up of clinical and voice-based data that was comprised of features including jitter, shimmer, RPDE, PPE, and HNR. The data was separated into testing and training sets to allow objective testing of the models. The choice of Python was made by the researchers due to comprehensive machine learning framework that offers rich tools and the programming language facilitates simple management of data and the language offers excellent data analysis of biomedical data. The process of implementation began by preprocessing data that was meant to increase the integrity and uniformity of the data. It was done by means of fill in missing data points, removal of redundant records, and removal of the background noise due to vocal characteristics, and then the Min-Max normalization was employed to reach to the normalized values of the features range. The process utilized feature importance by using Random Forest and Principal Component Analysis (PCA) to extract features which generated vital information concerning project specific attributes using feature extraction. These measures served to reduce the complexity in the computations without removing vital indicators that are required to classify the Parkinson disease accurately. Several machine learning models were constructed and evaluated using the processed dataset and they comprised the Support Vector Machine (SVM) model along with the Random Forest and K-Nearest Neighbors (KNN) and XGBoost models. Hyperparameter tuning was applied with cross-validation to help the team meet their performance objectives, as well as avoid overfitting their model. To ascertain model performance, the team assessed it using six measures that represented accuracy and precision and recall and F1-score and ROC-AUC and confusion matrices to obtain clinical reliability.

IV. EXPERIMENTATION EVALUATION

The suggested Parkinson's Disease detection system was tested to evaluate different machine learning models on the effectiveness of the system and its reliability in terms of operation. To conduct the impartial performance evaluation, the researchers separated the data of the clinical and voice-based feature into a training and testing set. The training sessions involved cross-validation of the system to be enhanced. The comparison made in the evaluation was on Support Vector Machine (SVM) Random Forest K-Nearest Neighbors (KNN) and XGBoost classifier to determine their efficiency in the detection of early Parkinson’s disease. In the study, the performance measured was using standard medical classification measures which covered accuracy precision recall F1-score ROC-AUC and analysis of a confusion matrix. Random Forest yielded productive outcomes on its full feature condition since it is capable of handling high-dimensional and inaccurate data. The results of SVM were better when it was applied with Principal Component Analysis (PCA) since it preserved high levels of classification

accuracy when it is tested on scaled-down sets of features. The overall accuracy and recall performance of KNN was highest when the data had balanced distribution that demonstrated collecting a lot of accuracy in identifying the cases of Parkinson, as compared to normal people. The results of the experiment indicate that one preprocessing method, as well as feature selection and dataset balance, influences the performance of the model. Analysis reveals that none of the models is better performing than other models since the results of models are determined by the data characteristics and step of preprocessing of data. The assessment reveals that machine learning techniques can be used to identify the Parkinson disease early and with high accuracy and certainty (Table I).

TABLE I. MODEL DESCRIPTION OF SYSTEMS

Sr. No.	Model	Description
1	Vector Machine Support (SVM)	By maximizing the margin between them, SVM generates an ideal hyperplane that divides data into classes. It works well with high-dimensional speech features used in Parkinson's disease identification and is efficient for non-linear data utilizing kernel functions.
2	KNNs, or K-Nearest Neighbors	KNN is a non-parametric technique that uses majority class of its closest neighbors to classify a data point. It is suitable for voice-based Parkinson detection classification since it is highly interpretable and performs well for small, balanced samples.
3	The Random Forest	An ensemble model that creates several decision trees and aggregate the results to increase accuracy. It works well with noisy data, avoids overfitting, and has high performance when numerous characteristics are present.
4	XGBoost	A gradient-boosted decision tree algorithm known for high accuracy, speed, and feature importance analysis. It sequentially corrects errors from previous trees and performs efficiently even with missing or noisy data.

V. FINDINGS AND INTERPRETATION

- The paper has shown that machine learning models can form a powerful system of diagnostic of the Parkinson disease that relies on clinical variables and speech patterns. The highest performance of the original model was achieved by the Random Forest classifier since it did not experience a drop in performance when operating with full 22 attributes and with mild data and embedded noise.
- The outcomes of the machine learning model indicate that various algorithms could be used to diagnose Parkinson disease in its early stages due to the identification functions they have. The KNN, Random Forest, and SVM models involve the clinical and speech information regarding an individual with or without Parkinson disease in their models, depending on their criteria of excellence.

VI. FUTURE ENHANCEMENTS

Parkinson disease detection system requires various approaches in order to improve the functions and increase the capabilities of the system (Table II). At the top of the list of the important advances, there will be the larger and more diverse datasets, not only that will bring the model closer to the actuality, but it will also result in greater adaptability to multiple age groups, disease phases, and recording types.

TABLE II. OVERVIEW OF ISSUES, APPROACHES AND METHODS

Sr. No.	Author(s) [Ref]	Issue Discussed	Approach & Method
1	Shrihari K. Kulkarni, K.R. Sumana [1]	Weakly regulated data collection; reliance on motion data	For PD prediction, ML and DL models (Decision Tree, Logistic Regression, Naïve Bayes, RNN) were compared.
2	Yatharth Nakul, Ankit Gupta, Hritik Sachdeva [2]	Need for higher accuracy; classification errors; slow output progression	Applied supervised ML (RF, SVM, Naïve Bayes, KNN), hyperparameter tuning, confusion matrix analysis
3	Wu Wang, Junho Lee, Fouzi Harrou, Ying Sun [3]	Need for robust early PD detection; low interpretability in DL	Used SGD, Feedforward Neural Network (FNN), LDA; achieved 96.45% accuracy
4	Muthumanickam S., Gayathri J., Eunice Daphne J. [4]	Difficulty interpreting deep models	Compared SVM, FBANN, Random Tree, Logistic Regression, LDA, CNN, DBN, DNN classifiers
5	Jayashree R. J., Ganesh S., Karanth S. C., Lalitha S. [6]	PD voice recognition in loud environments; restricted feature range	Extracted spectral & temporal features (STFT, Spectral Contrast, ZCR); used XGBoost, RF, Logistic Regression
6	Wroge et al. [5]	Need for voice-based telemonitoring	ML classification using voice biomarkers for PD detection

7	Arefi Shirvan & Tahami [10]	Difficulty selecting meaningful features for PD	Genetic Algorithm for feature selection + KNN classification
8	Shahbakhi, Far & Tahami [12]	Improving prediction accuracy in speech-based PD detection	Genetic Algorithm + SVM hybrid approach
9	Abhishek M. S. et al. [9]	Identifying PD from speech recordings	Machine learning classifiers applied to voice datasets
10	Alkhatib et al. [14]	Early detection of PD using gait abnormalities	Gait analysis and ML-based gait classification

VII. Conclusion

The project applies machine learning techniques to detect and diagnose an adapted version of the Parkinson disease among its early progression using both speech and clinical information. The system has obtained correct results in terms of the use of SVM KNN Random Forest and XGBoost classifiers to differentiate between the patterns of people with Parkinson and healthy people.

Appendix A System Architecture And Implementation Details

Machine learning algorithms are the preferred AI and machine learning solution that is used in the system. The program is based on the official programming language, framework (Django/Flask), which is used to detect Parkinson. There are two kinds of detection that the system carries out that comprise healthy human beings and parkinson. VS Code is the assessment tool applied in this research. The system is used in a test environment which can model a local network environment. The authors demonstrate their blind gratefulness to the Department of Computer science and engineering in the Anjuman College of Engineering, Nagpur as they have given them all the academic requirements to carry out this study.

References

[1] Shrihari K. Kulkarni, K. R. Sumana, Detection of Parkinson is a disease Using Machine Learning and Deep Learning Algorithms, International Journal of Engineering Science Invention (IJESI), Vol. 8, Issue 8, 2021.
[2] Yatharth Nakul, Ankit Gupta, Hritik Sachdeva, Parkinson Disease Detection Using Machine Learning Algorithms, International Journal of Science and Research (IJSR), Vol. 10, Issue 6, 2021.
[3] Wu Wang, Junho Lee, Fouzi Harrou, Ying Sun, Early Detection of Parkinson is a disease Using Deep Learning and Machine Learning, IEEE Access, Vol. 8, 2020.
[4] Muthumanickam S., Gayathri J., Eunice Daphne J., Parkinson is a disease Detection and Classification Using Machine Learning and Deep Learning Algorithms – A Survey, IJESI, Vol. 7, Issue 5, 2018.
[5] Timothy J. Wroge, Yasin Ozkanca, Cenk Demiroglu, Dong Si, David C. Atkins, Reza Hosseini Ghomi, Parkinson is a disease Diagnosis Using Machine Learning and Voice, IEEE SPMB, 2018.
[6] Jayashree R. J., Ganesh S., Karanth S.C., Lalitha S., Automatic Detection of Parkinson Speech Under Noisy Environment, Advances in Computing and Network Communications, Springer, 2021.
[7] Max A. Little, Patrick E. McSharry, Eric J. Hunter, Jennifer Spielman, Lorraine O. Ramig, Suitability of Dysphonia Measurements for Telemonitoring of Parkinson is a disease, IEEE Transactions on Biomedical Engineering, 2009.
[8] Jaichandran R., Leelavathy S., Usha Kiruthika S., Goutham Krishna, Mevin John Mathew, Jomon Baiju, Machine Learning Technique Based Parkinson is a disease Detection From Spiral and Voice Inputs, EJMC, 2020.
[9] Abhishek M. S., Chethan C. R., Aditya C. R., Divitha D., Nagaraju T. R., Diagnosis of Parkinson's Disorder Through Speech Data Using Machine Learning Algorithms, IJITEE, Vol. 9, Issue 3, 2020.
[10] R. Arefi Shirvan, E. Tahami, Voice Analysis for Detecting Parkinson is a disease Using Genetic Algorithm and KNN Classification Method, Iranian Conference on Biomedical Engineering, 2011.
[11] Gomathy C. K., Dheeraj Kumar Reddy B., Varsha B., Varshini B., The Parkinson is a disease Detection Using Machine Learning Techniques, IRJET, Vol. 8, Issue 10, 2021.
[12] Mohammad Shahbakhi, Danial Taheri Far, Ehsan Tahami, Speech Analysis for Diagnosis of Parkinson is a disease Using Genetic Algorithm and Support Vector Machine, Journal of Biomedical Science and Engineering, 2014.
[13] Made Satria Wibawa, Hanung Adi Nugroho, Noor Akhmad Setiawan, Performance Evaluation of Feature Selection and Classification Methods in Diagnosing Parkinson is a disease Based on Voice Features, ICSITECH, 2015.
[14] R. Alkhatib, M. O. Diab, C. Corbier, M. E. Badaoui, Machine Learning Algorithm for Gait Analysis and Classification on Early Detection of Parkinson, IEEE Sensors Letters, Vol. 4, No. 6, 2020.