EARLY PREDICTION OF PARKINSON USING MACHINE LEARNING

Mr.T.RAMALINGAM. M.E.,Assistant Professor, Department of Computer Science and Engineering,

Mr.M.SANJAY PRAKASH, B.E, Student of Computer Science Engineering

Mr.D.BARATHRAJ, B.E, Student of Computer science and Engineering

St. Joseph College of Engineering, Sriperumbudur, Chennai.

Abstract:

Parkinson's Disease (PD) is a progressive neurodegenerative disorder characterized by motor and non-motor symptoms, impacting the quality of life of affected individuals. Early diagnosis and prediction of PD are crucial for timely intervention and management, yet traditional diagnostic methods often lack sensitivity and specificity. Machine learning (ML) techniques have emerged as promising tools for early prediction and diagnosis of PD, leveraging various biomarkers and clinical data. This abstract presents a comprehensive review of recent advancements in ML-based approaches for early prediction of PD. Firstly, we provide an overview of PD pathophysiology and clinical manifestations, highlighting the importance of early detection. Next, we delve into the key biomarkers and data sources utilized in ML models for PD prediction, including genetic, imaging, and clinical features. Furthermore, we discuss the diverse ML algorithms employed in PD prediction, ranging from traditional classifiers to deep learning architectures. We analyze the strengths and limitations of each approach, emphasizing the need for robust feature selection, model interpretability, and validation on diverse datasets.

Moreover, we examine the integration of multimodal data and the potential of wearable sensors and mobile health technologies for continuous monitoring and early detection of PD-related symptoms. Additionally, we address the ethical considerations, challenges, and future directions in ML-based PD prediction, such as data privacy, model generalization, and clinical implementation.

In conclusion, ML holds immense promise for early prediction of Parkinson's Disease, offering personalized and timely interventions for improved patient outcomes. However, further research is needed to address existing limitations and facilitate the translation of ML models into clinical practice.

Introduction:

Parkinson's Disease (PD) is a complex and progressive neurodegenerative disorder that affects millions of individuals worldwide, with increasing prevalence due to aging populations. Characterized by a range of motor and non-motor symptoms, PD significantly impacts the quality of life of affected individuals and poses substantial challenges for healthcare systems globally. While there is currently no cure for PD, early diagnosis and intervention are crucial for symptom management, delaying disease progression, and improving patient outcomes.

Traditionally, the diagnosis of PD relies on clinical evaluation, often at a stage when significant neuronal damage has already occurred. Moreover, the clinical presentation of PD can vary widely among individuals, posing challenges for accurate diagnosis, particularly in the early stages when symptoms may be subtle or non-specific. As a result, there is a growing interest in leveraging advanced technologies, such as machine learning (ML), to facilitate early prediction and diagnosis of PD.

Machine learning techniques offer the potential to analyze large and heterogeneous datasets, including genetic, imaging, and clinical data, to identify patterns and biomarkers indicative of PD onset or progression. By integrating diverse data sources and employing sophisticated algorithms, ML models can provide personalized risk assessment and predictive analytics, enabling proactive interventions and personalized treatment strategies.

In recent years, there has been a proliferation of research focusing on ML-based approaches for early prediction of PD. These efforts have yielded promising results, demonstrating the feasibility of leveraging ML techniques to identify preclinical biomarkers, predict disease progression trajectories, and stratify individuals at risk of developing PD.

However, despite the growing interest and potential of ML in PD prediction, several challenges remain to be addressed. These include the need for large and representative datasets, robust feature selection methods, model interpretability, and validation on diverse populations. Moreover, ethical considerations, such as data privacy and equity in access to predictive technologies, must be carefully considered to ensure responsible and equitable deployment of ML models in clinical practice.

In this paper, we aim to provide a comprehensive overview of recent advancements in ML-based approaches for early prediction of Parkinson's Disease. We will review key biomarkers, data sources, and ML algorithms employed in PD prediction, as well as discuss challenges, ethical considerations, and future directions in this rapidly evolving field. By synthesizing current knowledge and identifying gaps in research, we hope to contribute to the development of effective and ethically sound strategies for early prediction and intervention in PD.

Literature Survey:

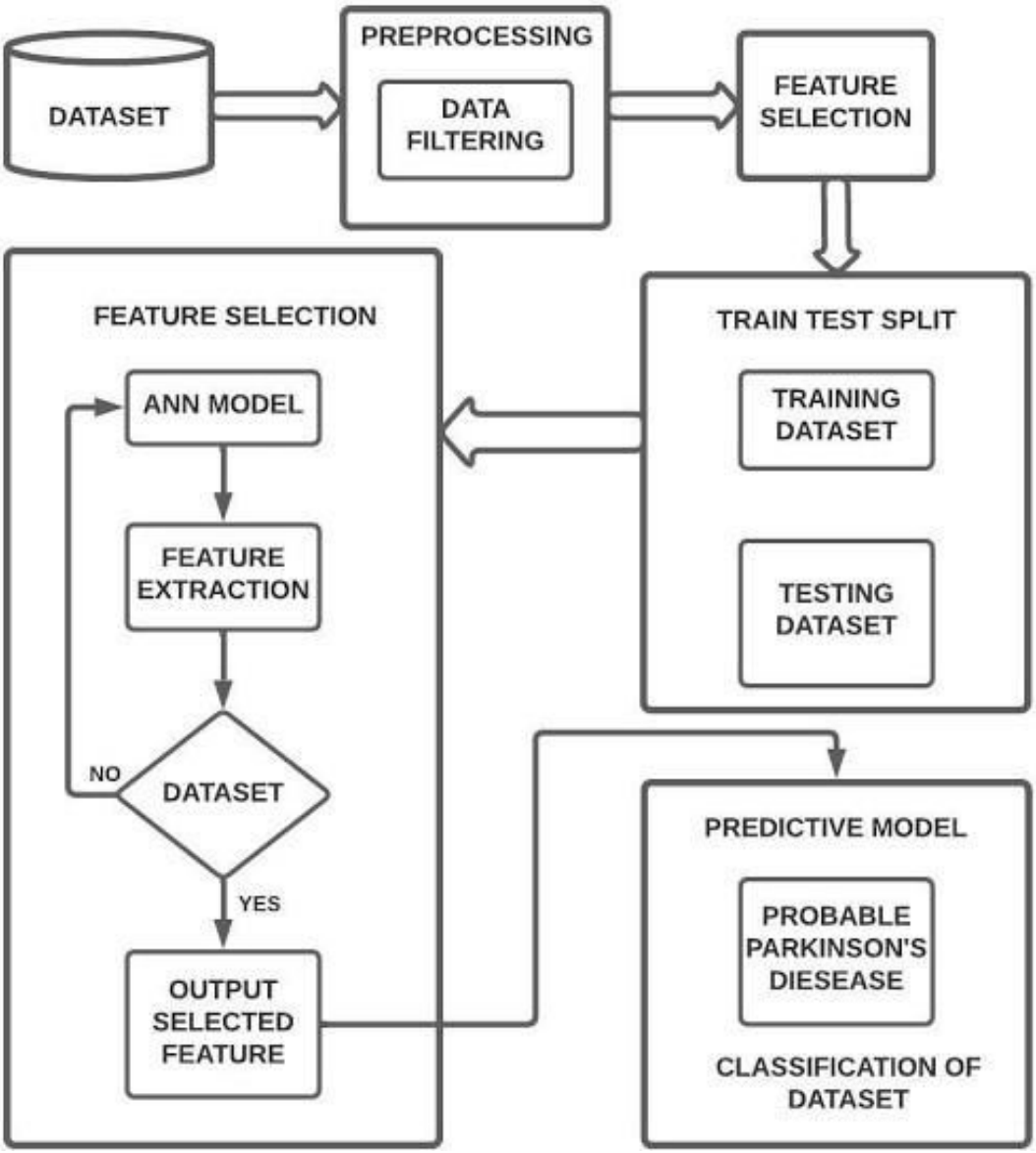
Previous studies to predict PD have been implemented on MRI scans, gait and genetic data, but research on audio impairment for early detection is minimal. For instance, Bilal et. al. [7] studied genetic data to predict the onset of PD in senior patients with SVM model. They trained an SVM model to reach an accuracy of 0.889, while this research paper describes an improved SVM model with an accuracy of 0.9183. These results also corroborate the
merits of classification of PD based on audio data, over genetic data. Raundale, Thosar and Rane [8] used keystroke data from UCI telemonitoring dataset to train a Random Forest classifier to predict the severity of PD in older patients. Cordella et. al. [9] use audio data to classify PWP, however their models are heavily reliant on MATLAB. Our research uses open-source models trained in Python, that are faster and memory efficient.
Majority of research done emphasizes the use of deep learning in PD detection, such as, Ali et. al. [10] who explain the use of ensemble deep learning models applied to phonation data, to predict the progress of Parkinson's disease. Their work lacked the use of feature selection that would improve Deep learning model (DNN)
performance. Hence, this paper implements PCA on 22 attributes to select 7 major voice modalities in PD detection. Huang et. al. [11] aim to reduce PD diagnosis dependence on wearable equipment by training a traditional decision tree on 12 complex speech features of the MDVR-KCL [12] dataset. Wodzinski et. al. [13] trained a ResNet model on images of audio data, instead of training the model on the nuances of the frequency of audio. Wroge et. al [14] aimed to remove subjectivity of doctors in prediction of PD using an unbiased ML model, however their results achieved peak accuracy of 85% only.
Wang et. al. [15] implemented 12 machine learning models on 401 voice biomarkers dataset to classify patients
as PD or not. They built a custom deep learning model (DEEP) with a classification accuracy of 96.45%, however
the model was expensive due to large memory requirements. Alkhatib et. al. [16] implemented a linear classification model with 95% accuracy to characterize shuffling movement of PD patients. Their study focused on gait of patient and future work encouraged the use of audio and sleep data to improve the results. Ricciardi et. al [17] performed spatial-temporal analysis of brain MRI scans. They implemented decision trees, random forest and KNN to detect Mild Cognitive Impairment (MCI) in PWP. However, dataset was small and artificial data augmentation [18] was needed. A. U. Haq and colleagues [19] implemented L1-support SVM, without feature identification on vowel phonation dataset for neurological disorder patients. Their paper focused on patient age group of 46-85 years,

PROPOSED METHODOLOGY:

The proposed methodology collects audio data from PPMI [21] and UCI about Parkinson's patients voice

- modulations. Dataset contains information about jitter, shimmer and MDVP of vowel phonations. Data is
- preprocessed, analyzed and visualized for a thorough understanding of the attributes. Four models – Logistic
- regression, SVM, Random Forest Regressor and K nearest neighbors – are trained on 75% of the data. Models are
- trained to classify given audio data into PD or healthy, based on variations in frequency. Models are tested on 25%

- of the data and evaluated based on sensitivity, precision, accuracy, confusion matrix [22] and ROC-AUC score.
- Figure 1 illustrates the generic process implemented. It demonstrates the stages of data ingestion from PPMI
- database, separation of data into testing and training sets, training of four models on data and validation of results.



- This research paper aims to identify the most relevant attributes in classification of PD and impact of imbalance
- in medical data in classification. Keeping in mind these requirements, 3 approaches have been implemented –
- Training on complete dataset that serves as a baseline test for PD classification, training on PCA identified attributes
- and training on 109 records obtained after dataset balancing. The algorithms used in each approach are described
- below:
- Algorithm for approach 1: Models are trained on 22 attributes of data
- • Collect MDVP audio data from PPPMI and UCI databases
- • Perform data analysis to detect skew, imbalance and distribution of variables in data
- • Scale the data to common range using Standard Scaler
- • Split dataset into testing and training sets, where training data is 75% of total
- • Train SVM, logistic regression, random forest and KNN models.
- Algorithm for approach 2: Principal Component Analysis (PCA) is applied to identify 5 key attributes
- • Collect MDVP audio data from PPPMI and UCI databases
- • Perform data analysis to detect skew, imbalance and distribution of variables in data
- • Scale the data to a common range using Standard Scaler
- • Identify variance in every column of data and apply Principal Component Analysis (PCA) to identify 5
- most relevant features to model training, out of 22 attributes.
- • Split dataset into testing and training sets, where training data is 75% of total
- • Retrain SVM, logistic regression, random forest and KNN models.
- • Compare classification results using confusion matrix, ROC-AUC curve and accuracy
- Algorithm for approach 3: Imbalance removal in dataset
- • Collect MDVP audio data from PPPMI and UCI databases
- • Perform data analysis to detect skew, imbalance and distribution of variables in data
- • The dataset is imbalanced, with 109 records of PWP and 40 records of normal people, as illustrated in

- figure 2(a). The imbalance is resolved by up sampling [23] the minority class to reach 109 records each,
- as illustrated in figure 2(b).
- • Scale the data to common range using Standard Scaler
- • Split dataset into testing and training sets, where training data is 75% of total
- • Retrain SVM, logistic regression, random forest and KNN models.
- • Compare classification results using confusion matrix, ROC-AUC curve and accuracy
- Model Deployment:
- Integrate the trained models into the credit card transaction processing pipeline for automated risk assessment.
- Implement APIs or microservices to facilitate seamless communication between the model and transactional systems.
- Ensure model interpretability and explainability to enable stakeholders to understand model decisions and predictions.

CONCLUSION:

Parkinson's disease classification using vowel phonation data gives an 91.835% accuracy and 0.95 sensitivity for Random Forest classifier. Results of the Random Forest model are ideal, due to equal importance given to all 22 attributes in MDVP dataset. This paper also highlights the results of the SVM model that gives an accuracy of
91.836% and sensitivity of 0.94, after PCA is applied to the dataset. Both SVM and Random Forest models perform well for outliers and are robust models. The models predict no false positives in the results. K nearest neighbor (KNN) model also performs well for balanced dataset, as classification into 2 categories without presumptions of
data is favored. Thus, we recommend the use of Random Forest model to classify progress of the disease. It is a non- invasive, simple and accurate technique to provide long-term relief to PWP, globally.

In the future, we propose to use audio and REM sleep data to improve the results, as audio data alone is not a sufficient biomarker for classification of Parkinson's disease. We hope these findings encourage the use of mobile recorded audio to classify PD through telemedicine.

REFERENCES:

[1] Prabhavathi, K., Patil, S. (2022). "Tremors and Bradykinesia. In: Arjunan, S.P., Kumar, D.K. (eds) Techniques for Assessment of Parkinsonism for Diagnosis https://doi.org/10.1007/978-981-16-3056-9_9

[2] Braak, H., Braak, E. (2000) "Pathoanatomy of Parkinson's disease" J Neurol 247, II3–II10. https://doi.org/10.1007/PL00007758

[3] F. Amato, I. Rechichi, L. Borzì and G. Olmo, (2022), "Sleep Quality through Vocal Analysis: A Telemedicine Application," 2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), 706-

711, doi: 10.1109/PerComWorkshops53856.2022.9767372.

[4] Neighbors C, Song SA. “Dysphonia” (2022) StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing.

[5] Serge Pinto, Canan Ozsancak, Elina Tripoliti, Stéphane Thobois, Patricia Limousin - Dowsey, Pascal Auzou, “Treatments for dysarthria in Parkinson's disease”, (2004) The Lancet Neurology, 3(9): 547-556, ISSN 1474-4422, https://doi.org/10.1016/S1474-4422(04)00854-3.

[6] Nicoló G. Pozzi, Ioannis U. Isaias (2022), “Chapter 19 - Adaptive deep brain stimulation: Retuning Parkinson's disease”, Elsevier 184: 273-

AUTHOR 1

Mr..T.RAMALINGAM M.E., Assistant Professor, Department of Computer Science and Engineering at St.Joseph College of Engineering, Sriperumbudur, Chennai, TamilNadu.

AUTHOR 2

Mr.M.SANJAY PRAKASH B.E., Student of Computer Science and Engineering at St.Joseph College of Engineering, Sriperumbudur, Chennai, TamilNadu. I had attended many Workshops, Seminars in Python, Machine Learning and Data Analytics.

AUTHOR 3

Mr.D.BARATHRAJ B.E., Student of Computer Science and Engineering at St.Joseph College of Engineering, Sriperumbudur, Chennai, Tamil Nadu. I had attended many Workshops and Seminars in the area of Python and Machine Learning.