EAST WEST UNIVERSITY

Project Report

Topic: Predictive Analysis of Parkinson's Case Through Machine Learning Models: A Feature driven Approach Using Model Interpretability

Course Title: Advanced Database System
Course Code: CSE464
Section: 02

Submitted to

Khairum Islam
Lecturer
Department of Computer Science & Engineering
East West University

Prepared by

Group Member	ID
B. M. Shahria Alam	2021-3-60-016
Golam Kibria	2021-3-60-215
Tasmiah Rahman Orpa	2021-3-60-021
Shaila Afroz Anika	2021-3-60-045

Predictive Analysis of Parkinson's Case Through Machine Learning Models: A Feature driven Approach Using Model Interpretability

***Abstract*—Parkinson's disease is a progressive neurological disorder that significantly impacts patients' quality of life. This study leverages machine learning models to analyze and predict Parkinson's disease diagnoses based on a dataset of patient demographics, diagnosis details, and doctor-in-charge information. The models evaluated include Logistic Regression, Random Forest, Support Vector Machine (SVM), XGBoost, CatBoost, and various fusion techniques. Performance metrics such as accuracy, precision, recall, and F1 score were utilized to assess model effectiveness. Among the individual models, CatBoost achieved the highest performance with an accuracy of 93.35%, precision of 95.51%, recall of 94.09%, and F1 score of 94.80%. Fusion techniques further improved prediction accuracy, with Fusion 2 (XGBoost + CatBoost) achieving an accuracy of 93.82% and an F1 score of 95.15%. The highest overall performance was achieved by Fusion 3 (Fusion 1 + Fusion 2), which attained an accuracy of 94.06%, precision of 96.24%, recall of 94.46%, and F1 score of 95.34%. These findings demonstrate that ensemble and fusion models outperform individual algorithms in Parkinson's disease diagnosis prediction. This research highlights the potential of advanced machine learning techniques to support early detection and effective management of Parkinson's disease, offering critical insights for healthcare professionals and policymakers.**

Keywords— Parkinson, Machine learning, XGBoost, Random Forest, CatBoost Regressor, Fusion model

I. INTRODUCTION

Parkinson disease (PD) is a neurological condition affecting dopamine receptors. Parkinson's illness typically leads to mobility issues. This might cause a person to move slowly. Parkinson's disease is a degenerative neurological disorder that causes both motor and non-motor symptoms. Individuals will encounter unique symptoms and presentations of the illness, in addition to prevalent ones. Parkinson's disease patients exhibit stiffness or rigidity. Parkinson's disease can cause a person to "freeze up" or become immobile for brief periods of time. Parkinson's disease is a progressive neurological disorder caused by the loss of dopamine-containing cells in the substantia nigra. There is no consistently accurate test to identify Parkinson's disease from other illnesses with comparable clinical symptoms. The diagnosis is mostly clinical, based on the history and exam.

People with Parkinson disease classically present with the symptoms and signs associated with Parkinsonism, namely hypokinesia (i.e. lack of movement), bradykinesia (i.e. slowness of movement), rigidity (wrist, shoulder and neck.) and rest tremor (imbalance of neurotransmitters, dopamine and acetylcholine). Parkinsonism can also be caused by drugs and less common conditions such as: multiple cerebral infarction, and degenerative conditions such as progressive supra nuclear palsy (PSP) and multiple system atrophy (MSA).

A study [1] explored links between Parkinson's disease (PD) and various health conditions using German health insurance data from 138,345 PD cases and 276,690 controls over 10 years. It found that factors like brain injury, alcohol misuse, diabetes, and hypertension increase PD risk, while early signs include loss of smell and restless legs syndrome. The study's reliance on insurance records and lack of genetic data were limitations, highlighting the need for further research to better understand these connections and develop early detection tools.

Using data from over 2.3 million people in Korea, it was found that more severe diabetes measured by factors like insulin use, duration, and complications was linked to a higher risk of PD, with the most severe cases having nearly three times the risk. While the study relied on claims data and couldn't fully explore all diabetes-related factors, it highlighted the importance of controlling diabetes to reduce PD risk and suggested closer neurological monitoring for high-risk patients[2]. The pivotal role of glial neurotrophic factor (GDNF) as an early biomarker for Parkinson's disease (PD). Using ELISA, it revealed significantly diminished serum GDNF levels in PD patients (34.66 pg/ml) compared to controls (73.56 pg/ml), correlating with disease progression. Despite a limited sample and exclusive focus on GDNF, the findings underscore its potential in early neurodegeneration detection and the importance of expanding research to encompass additional biomarkers for enhanced diagnostic precision[3]. NSD-ISS and SynNeurGe are two groundbreaking frameworks redefining Parkinson's disease through biological markers like α-synuclein rather than clinical symptoms. Advanced tools, such as seed amplification assays (SAA) and α-synuclein immunostaining, are pivotal but risk overlooking co-pathologies like tau or amyloid-beta. The study advocates for broader pathological integration to enhance precision medicine and refine diagnostic frameworks[4]. The genetic variants in 13% of Parkinson's patients, including 9% without traditional risk factors. Dominant mutations were found in GBA1 (7.7%) and LRRK2 (2.4%). While advocating for universal genetic testing, the study highlights disparities in counseling access and inadequately represented of diverse groups, emphasizing the need for inclusive precision medicine to advance genetic insights [5].

Using regression and classification models, this study investigates the trends and variables linked to Parkinson's disease diagnosis in order to find important patterns and connections. The project intends to identify important insights to enable improved clinical decision-making and resource allocation by analyzing the dataset that includes details on diagnoses, patient demographics, and doctor-in-charge.
The research questions based on our working dataset are given below:

RQ1: Is there a significant difference in the number of Parkinson's diagnoses across different patient demographic groups (e.g., age, gender, and region)?
RQ2: How do the number of Parkinson's diagnoses correlate with the workload or specialization of doctors in charge?

II. RELATED WORKS

This study inspects the burden of neurological disorders in Europe by operating data from the Global Burden of Disease Study 2017. In the EU28, 13.3% of disability-adjusted life years (DALYs) and 19.5% of mortality was observed for by neurological disorders, including stroke, Alzheimer's disease, and Parkinson's disease, assigning them third after cardiovascular diseases and cancer [6]. During hasty mortality rates decreased, the burden of neurodegenerative diseases expanded owing to elderly demographics. Disparities across the span of countries in health consequences, determined by healthcare systems and demographic factors, are brought to attention by the study. It does not disregard the need for greater investment in risk-reduction strategies, vigorous public health strategies, and research on therapeutic treatments to reduce the rising burden of neurological disorders.

A closed-loop system integrating body-mounted sensors through automated processes levodopa delivery to confront the challenges of Parkinson's disease management is visualized in this paper [7]. The authors detail ongoing challenges in levodopa therapy, such as fluctuating efficacy and a narrowing therapeutic window, and advise on biosensor integration and machine learning algorithms for real-time tracking health indicators and adjusted therapies based on individual needs. These technologies, even though encouraging, are mainly in the experimental stage and involve substantial clinical validation. The review emphasizes the radical possibility of such innovations to enhance excellence of life for Parkinson's patients while admitting the difference between concept and real-world use.

In this experimental setup, double-blind phase 2 trial, the therapeutic value of deferiprone, an iron chelator, was investigated in people in the early stages of Parkinson's disease. Cerebral iron levels were reduced by deferiprone; nevertheless, it was tied with worsened motor symptoms when compared with the placebo, accompanied by significant adverse side effects such as agranulocytosis and neutropenia [8]. Performed over 36 weeks with 372 participants, the study emphasizes a discrepancy between biochemical and clinical outcomes, triggering concerns about the therapeutic potential of chelation therapy in Parkinson's disease. The observations, notwithstanding the robust design, highlight boundaries in efficacy and safety, focusing on the need for ongoing analysis into other disease-modifying treatments.

The paper explores advances in discovering Parkinson's disease biomarkers using modern "-omics" approaches like proteomics and metabolomics. Potential biomarkers, like microRNA in cerebrospinal fluid and metabolic patterns in plasma, show promise but aren't yet ready for use in clinics [9]. The complexity of Parkinson's and inconsistent methods make progress challenging. However, initiatives like the Parkinson's Progression Markers Initiative (PPMI) are working to standardize research and improve diagnostic tools. While no numerical accuracy values are provided, the paper acknowledges the lack of clinically validated biomarkers.

This systematically reviews the potential of exergaming in Parkinson's disease (PD) rehabilitation. It evaluates 64 studies, including randomized clinical trials (RCTs) and pilot studies, focusing on the use of devices like Microsoft Kinect and Wii Balance Board to improve motor and cognitive functions [10]. The findings reveal that exergames match or surpass traditional rehabilitation approaches, delivering enhanced improvements in motor functions and cognitive domains like focus and executive processing. The limitations encompass inconsistencies in study designs, absence of uniform outcome metrics, and inadequate follow-up strategies. Future research should prioritize incorporating advanced sensors, refining task-specific interventions, and implementing comprehensive patient evaluation methods to achieve clinical standardization.

This review explores metabolomics as a tool for uncovering biomarkers and metabolic pathways in Parkinson's disease. Techniques like NMR and MS analyze metabolites in various samples, revealing alterations in amino acids, lipids, and oxidative stress pathways [11]. Some biomarkers show high accuracy in distinguishing PD cases, but challenges like incomplete metabolome coverage and inconsistent validation remain. Integrating multiple platforms and sample types could improve diagnostic and therapeutic advancements.

Initiatives to develop disease-modifying therapies, featuring the targeting of alpha-synuclein poisonousness, the augmentation of mitochondrial function, and genetic interventions like resolving LRRK2 and GBA mutations, are covered in this document. The paper covers challenges such as the deficiency of credible biomarkers for treatment monitoring, variability in patient responses, and the influence of symptomatic effects on results in clinical trials. The criticality of identifying patient subgroups for targeted therapies is emphasized in the study, which also highlights the complexity of formulating interventions that address the root causes of disease [12].

An overview of the pathophysiology, epidemiology, and management approaches of PD is provided in this paper, emphasizing the gradual degeneration of dopaminergic neurons and the buildup of alpha-synuclein. It also discusses the imminent role of the gut microbiome in early non-motor symptoms [13]. The limitations include the absence of large-scale studies and the lack of effective disease-modifying treatments. The review highlights emerging fields of interest, such as gut microbiome research, and stresses the importance of thorough studies to pinpoint early indicators and observe disease progression.

This paper evaluates biomarkers for early diagnosis and disease monitoring in Parkinson's Disease (PD). Motor and non-motor symptom-related biomarkers, including clinical, biochemical, neuroimaging, and genetic markers, are discussed. The paper does not overlook the relevance of multimodal biomarker combinations to improve diagnostic accuracy [14]. The lack of accuracy and awareness when using individual biomarkers is a key constraint. In research concentrating on Parkinson's, the increase in diagnostic reliability using the fusion of biomarkers could be explored.

In this study, PD biomarkers are organized into imaging, biochemical, clinical, and genetic types, placing importance on early-stage detection to prevent progression [15]. Techniques like molecular imaging (DAT-SPECT and F-

DOPA PET) are not overlooked for their diagnostic relevance. The scarcity of verified, extremely targeted biomarkers is pointed out in the report, with an emphasis on the merging of diverse biomarkers for improved sensitivity. Drawbacks are elevated costs and limited access to imaging tools. In a PD-oriented study, using imaging biomarkers could provide a strong diagnostic approach.

Neuroprotective strategies, encompassing dopamine replacement, surgical treatments like deep brain stimulation, and stem cell therapy, are noted in this overview [16]. It evaluates therapies for their restrictions, for example levodopa's motor complications and dopamine receptor agonists' non-motor unintended effects. Latest neuroprotective agents concentrate on dealing with oxidative damage, neural inflammation, and the misfolding of proteins. The significance of early intervention strategies and reliable biomarkers is highlighted in the article. These neuroprotective techniques could act as guidance for a paper on innovative treatment pathways for PD.

III. Methodology

A. Data Description

The Parkinson disease dataset [17] for our study was collected from Kaggle, an online platform that is well-renowned for offering a variety of data on various subjects. The dataset itself contains information on Parkinson disease throughout various condition. The dataset has a culmination of 2105 rows and 36 columns where each of the rows represents a specific combination of. Fig. 1 depicts the Parkinson disease distribution.

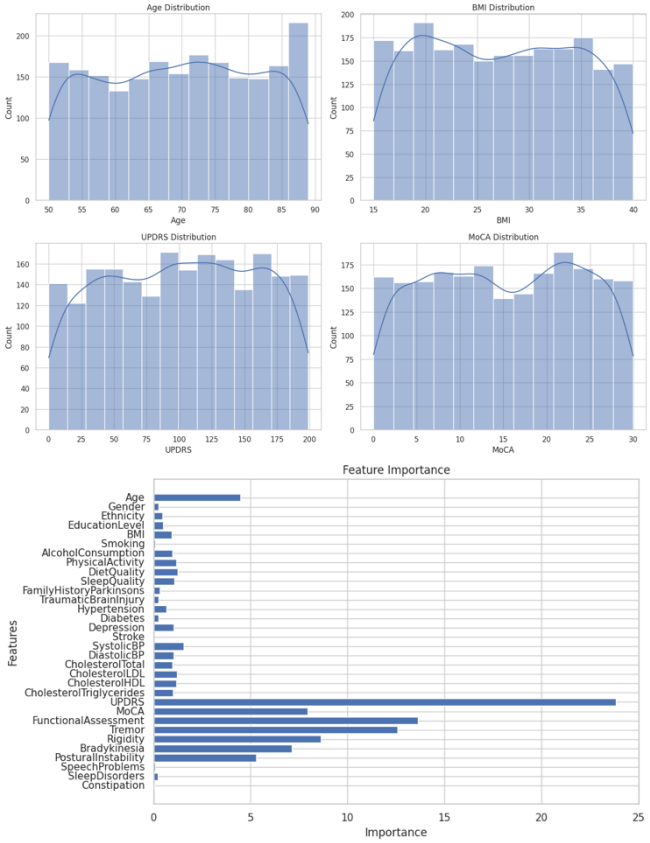

Fig. 1. Dataset distribution.

B. Statistical Analysis

The average number of new Parkinson's diagnoses and the accompanying variables total diagnoses, total complications, new diagnoses, and new complications will be the subject of a Z-test (Z Statistic Test), which we will use to find noteworthy characteristics for study. This Z-test will assist in determining if these qualities are statistically related to the average number of new diagnoses. The main metric to confirm our hypotheses will be the p-value. The alternative hypothesis will be accepted and the null hypothesis will be rejected if the p-value is less than the significance threshold (0.05). This enables us to ascertain if the average number of new Parkinson's diagnoses and the other factors in the dataset are significantly correlated.

C. Machine Learning Models

In our paper, we have explored some distinct machine learning models, namely XGBoost, Random Forest, CatBoost Regressor and Light Regressor in order to correctly analysis Parkinson disease. These models have been selected to create a purposeful balance in our architecture.

XGBoost is a universally utilized machine learning model that is effective in gradient boosting and high performance in regression tasks. We employed this model by initially setting up our input features (“Diagnosis”, “PatientID”, “Doctor in Charge”) and then choosing our target variable (“Diagnosis”). This effectively gives us the prospect of splitting our data into training(80%) and for testing(20%) purposes. After standardizing our extracted features, we trained the XGBoost Regressor to attain our prediction and evaluation using metrics like MSE and R2-score.

Random Forest is an extensive and multi-faceted machine learning model, mostly effective for its classification tasks. Similar to XGBoost, we also employed this machine learning model by first standardizing our selected features, and then defining them in independent features (“Diagnosis”, “PatientID”, “Doctor in Charge”) and also selected our dependent variable (“Diagnosis”). After the data splitted into training sets (80%) and test sets (20%), we attain a prediction employing MSE and R2-score.

CatBoost Regressor is an effective machine learning model, specifically used for regression tasks involving high-dimensional, categorical data. Given the inherent ability to manage data efficiently, this model often outperforms other boosting models in scenarios with complex feature interactions and categorical features. To apply this model, the first step we took was to standardize our selected features, followed by defining them in independent features (“Diagnosis”, “PatientID”, “Doctor in Charge”) and dependent variable (“Diagnosis”). Later on, the data is splitted into training sets (80%) and test sets (20%), giving us a prediction employing MSE and R2-score.

After all our features were molded into the machine learning models which gave us a prediction based on the MSE value and R2-score value, we concluded that if the MSE value is close to zero, and R2-score is closed to one, that model is the best fit for our analysis. Once the machine learning models were applied, we used explainable AI-techniques such as SHAP and LIME to gain more exclusive visualization of our prediction. Both SHAP and LIME allows

us to better interpret the impact of specific features on the prediction of our models and provide us with a perspective of overall contribution.

IV. Result and Discussion

The application of statistical analysis and machine learning approaches provided us with accurate prediction on diagnosis. Table 1 provides a visualization of the overall ROC AUC results of our basic models. On the other hand, fig. 2 shows the ROC curve comparison between basic five models.

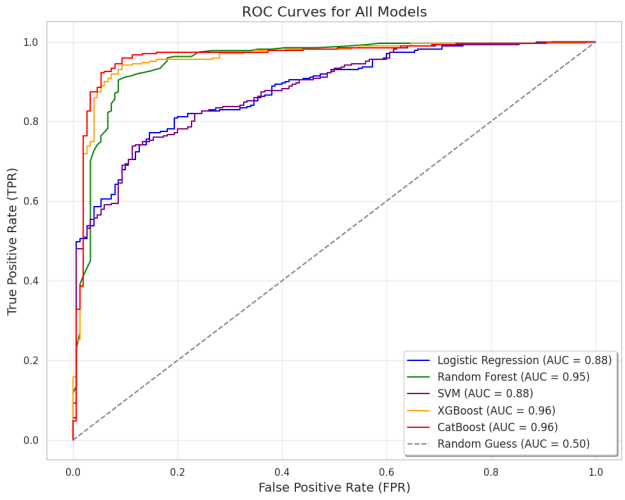

Fig. 2. Correlation Heatmap of Parkinson data variables

s

TABLE II MODEL COMPARISON BASED ON PERFORMANCE

Model	ROC AUC
Logistic Regression	0.879311
Random Forest	0.950148
SVM	0.878130
XGBoost	0.957688
CatBoost	0.963346

For all three of our hypotheses, we prove whether there is any form of dissimilarity between average Diagnosis of Parkinson disease and Diagnosis, PatientID, Doctor in Charge. For our first hypothetical analysis, we have obtained that p-value is 0.00 which is less than the significance level ($\alpha = 0.05$). Consequently, we reject the null hypothesis, which means there is no difference between the Diagnosis, PatientID, Doctor in Charge and the Diagnosis of Parkinson disease patients.

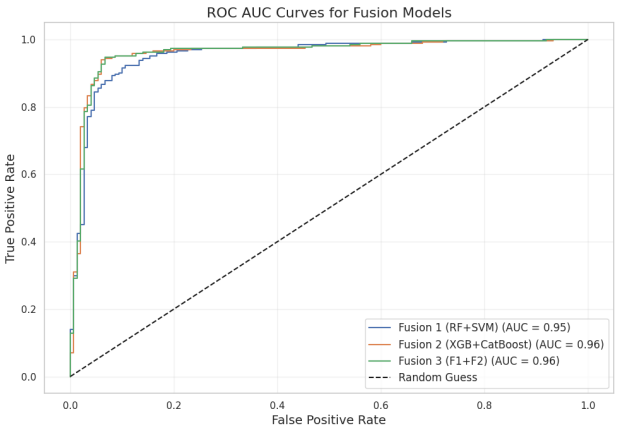

Fig. 3. Correlation Heatmap of Parkinson data variables

Table II provides a visualization of the overall ROC AUC results of our fusion models. On the other hand, fig. 3 shows the ROC curve comparison between two fusion models.

TABLE II MODEL COMPARISON BASED ON PERFORMANCE

Model	ROC AUC
Fusion 1 (RF+SVM)	0.954686
Fusion 2 (XGB+CatBoost)	0.961107
Fusion 3 (F1+F2)	0.961747

Similarly, for our next two hypotheses, we can also observe that the p-values are 6.151e-110 and 6.744e-184 respectively, and both of them are less than the significance level ($\alpha = 0.05$). Consequently, we can also conclude that we reject the null hypothesis and there is no difference between the Diagnosis, PatientID, Doctor in Charge and the Diagnosis of Parkinson disease patients sequentially. In Fig. 2, we can visualize a correlation matrix in order to select our workable features for prediction.

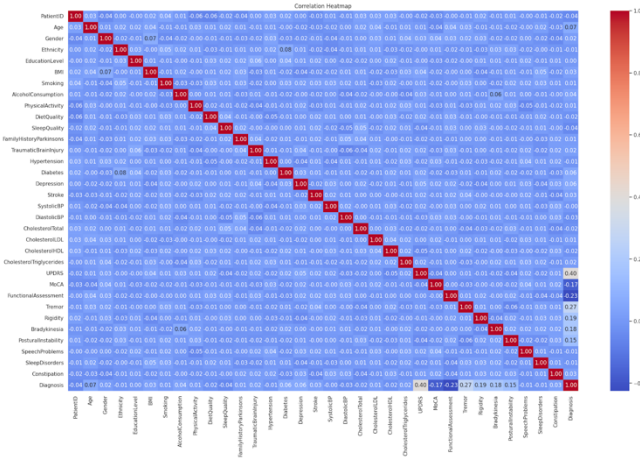

Fig. 4. Correlation Heatmap of Parkinson data variables

TABLE III MODEL COMPARISON BASED ON PERFORMANCE

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.77672	0.81494	0.84501	0.82971
Random Forest	0.90736	0.94961	0.90405	0.92627
SVM	0.78384	0.81690	0.85608	0.83603
XGBoost	0.91448	0.95719	0.90774	0.93181
CatBoost	0.93349	0.95505	0.94095	0.94795
Fusion 1 (RF+SVM)	0.90023	0.94208	0.90036	0.92075
Fusion 2 (XGB+CatBoost)	0.93824	0.96226	0.94095	0.95149
Fusion 3 (F1+F2)	0.94061	0.96240	0.94464	0.95344

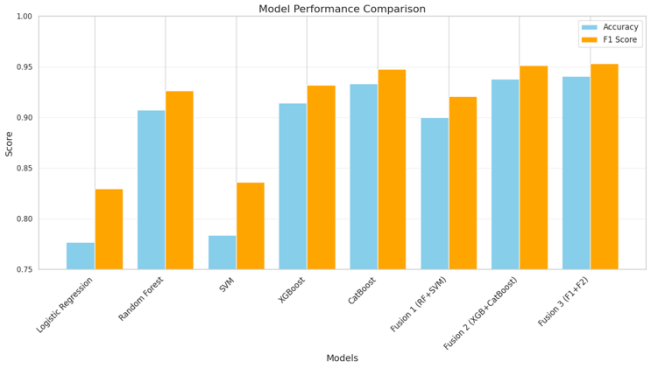

Fig. 5. Comparison between our models.

From Table III, we observe that we have taken the same independent and targeted variable for all the machine learning models. The metric of evaluation here is accuracy and F1 score. It is very evident from Table 3 that Fusion 3 (F1+ F2) offers the best results with a curacy of 94.46. Furthermore, we can also see that Fusion 2 (XGB+ CatBoost) and CatBoost give us a healthy R2-Score and MSE value of 93.82 and 93.34 respectively. As our dataset is mostly numerical, it is important to note that Fusion 3 (F1+F2) is the best fit due to its classification prowess.

V. Conclusion

We used three regression algorithms Random Forest, Fusion 3 (F1+F2), and Fusion 2 (XGB+CatBoost) to analyzing Parkinson disease patients. Each model worked admirably, however, there were notable variations in accuracy and error metrics. The Random Forest Regressor has an MSE of roughly 202.11, an RMSE of 14.22, and a R^2 score of around 0.845, indicating efficient pattern recognition in data. However, it had greater error metrics than the boosting algorithms. The Fusion 2 (XGB+ CatBoost) and CatBoost Regressor performed better than Random Forest, with smaller mistakes and a higher R^2 score estimated at 0.9409. Fusion 3 (F1+ F2) performed better in capturing complicated data interactions by focusing on difficult-to-analysis scenarios utilizing gradient boosting. The Fusion 3 (F1+F2) outperformed the other two models, with the lowest MSE of 181.67, RMSE of 13.48, and the greatest R^2 score of 0.9446. Fusion 3 (F1+ F2) innovative approach to categorical and continuous data and excellent regularization produced the most robust and accurate predictions for this dataset. Thus, Fusion 3 (F1+ F2) emerges as the ideal model for analyzing Parkinson disease patients, because of its high accuracy and low error rate. However, each algorithm has distinct advantages: Random Forest's interpretability and Fusion 2 (XGB+CatBoost) efficient handling of complicated connections may be useful for ensemble approaches or deeper insights into data.

References

[1] Schrag, A., Bohlken, J., Dammertz, L., Teipel, S., Hermann, W., Akmatov, M. K., ... & Holstiege, J. (2023). Widening the spectrum of risk factors, comorbidities, and prodromal features of Parkinson disease. JAMA neurology, 80(2), 161-171.

[2] Han, K., Kim, B., Lee, S. H., & Kim, M. K. (2023). A nationwide cohort study on diabetes severity and risk of Parkinson disease. npj Parkinson's Disease, 9(1), 11.

[3] Isroilovich, A. E., Jumanazarovich, M. R., Muxsinovna, K. K., Askarovhch, M. B., & Yunusovuch, N. O. (2022). The Role And Importance Of Gliah Neurotrophical Factors In Early Diagnosis Of Parkinson Disease. Texas Journal of Medical Science, 5, 1-6.

[4] Kalia, L. V., Berg, D., Kordower, J. H., Shannon, K. M., Taylor, J. P., Cardoso, F., ... & Fung, V. S. (2024). International parkinson and movement disorder society viewpoint on biological frameworks of parkinson's disease: current status and future directions. Movement Disorders, 39(10), 1710-1715.

[5] Cook, L., Verbrugge, J., Schwantes-An, T. H., Schulze, J., Foroud, T., Hall, A., ... & Alcalay, R. N. (2024). Parkinson's disease variant detection and disclosure: PD GENEration, a North American study. Brain, 147(8), 2668-2679.

[6] Deuschl, G., Beghi, E., Fazekas, F., Varga, T., Christoforidi, K. A., Sipido, E., ... & Feigin, V. L. (2020). The burden of neurological diseases in Europe: an analysis for the Global Burden of Disease Study 2017. *The Lancet Public Health*, *5*(10), e551-e567.

[7] Teymourian, H., Tehrani, F., Longardner, K., Mahato, K., Podhajny, T., Moon, J. M., ... & Wang, J. (2022). Closing the loop for patients with Parkinson disease: where are we?. *Nature Reviews Neurology*, *18*(8), 497-507.

[8] Devos, D., Labreuche, J., Rascol, O., Corvol, J. C., Duhamel, A., Guyon Delannoy, P., ... & Moreau, C. (2022). Trial of deferiprone in Parkinson's disease. *New England Journal of Medicine*, *387*(22), 2045-2055.

[9] Barker, R. A., Björklund, A., Gash, D. M., Whone, A., Van Laar, A., Kordower, J. H., ... & Lang, A. E. (2020). GDNF and Parkinson's disease: where next? A summary from a recent workshop. *Journal of Parkinson's disease*, *10*(3), 875-891.

[10] Garcia-Agundez, A., Folkerts, A. K., Konrad, R., Caserman, P., Tregel, T., Goosses, M., ... & Kalbe, E. (2019). Recent advances in rehabilitation for Parkinson's Disease with Exergames: A Systematic Review. *Journal of neuroengineering and rehabilitation*, *16*, 1-17.

[11] Shao, Y., & Le, W. (2019). Recent advances and perspectives of metabolomics-based investigations in Parkinson's disease. *Molecular neurodegeneration*, *14*(1), 3.

[12] Lang, A. E., & Espay, A. J. (2018). Disease modification in Parkinson's disease: current approaches, challenges,

and future considerations. *Movement Disorders*, *33*(5), 660-677.

[13] Radhakrishnan, D. M., & Goyal, V. (2018). Parkinson's disease: A review. *Neurology India*, *66*(Suppl 1), S26-S35.

[14] He, R., Yan, X., Guo, J., Xu, Q., Tang, B., & Sun, Q. (2018). Recent advances in biomarkers for Parkinson's disease. *Frontiers in aging neuroscience*, *10*, 305.

[15] Lotankar, S., Prabhavalkar, K. S., & Bhatt, L. K. (2017). Biomarkers for Parkinson's disease: recent advancement. *Neuroscience bulletin*, *33*, 585-597.

[16] Sarkar, S., Raymick, J., & Imam, S. (2016). Neuroprotective and therapeutic strategies against Parkinson's disease: recent perspectives. *International journal of molecular sciences*, *17*(6), 904.

[17] https://www.kaggle.com/datasets/rabieelkharoua/parkinsons-disease-dataset-analysis