Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Early Detection of Parkinson's Disease from Sleep Efficiency using Optimized Machine Learning Models

N Sai Keerthi [1], G Krishna Chaitanya[2]
[1]*M. Tech Candidate, Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, 522302, Andhra Pradesh, India.*
[2]*Associate Professor, Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, 522302, Andhra Pradesh, India.*

ARTICLE INFO

Received: 30 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Parkinson's disease (PD) a type of neurodegenerative disorder that affects elderly people. Patients with Parkinson's disease mostly have symptoms of muscle rigidity, speech difficulty, movement challenges. PD when not detected at the beginning may cause severe health issues, thus premature detection of PD is crucial and more challenging. The major early symptom for PD is sleep disorder, frequent awakening in the night, this leads them to lesser the quality of sleep, resulting in disease worsening as the time increases. With the sleep efficiency the PD can be detected at early stages when checked with other diagnosing methods including gait recognition, speech tremor data. PD can be treated through early detection; this enables patients to lead a normal life with proper medication. The rise of an aging population around the world is the major cause for the disease to be identified premature and accurately. This project, sleep efficiency data is used to identify Parkinson's disease. Machine learning models like AdaBoost (ADB), Extreme Gradient Boosting (XGB) and Extra Trees (ET) for the prediction of Parkinson's disease from sleep efficiency. Experimental results demonstrated that the AdaBoost model showed the highest accuracy of Parkinson's disease classification around 99.11%.

Keywords: Parkinson's Disease, Machine Learning, Boosting techniques, Extreme Gradient Boosting, Extra Trees models.

INTRODUCTION

Parkinson’s disease (PD) is a neurological condition that causes movement disorder use to the degradation of nerve cells, this is the condition caused when the dopamine production levels in the brain is reduced. According to the World Health Organization (WHO) reports 2023 [1], the disability and death of PD patients are dramatically increasing over the world. There are around 8.5 million patients affected due to PD around the world during 2019, this amount is elevated to 81% since the year 2000 and the total deaths increased 100% since 2000. There are therapies and medications available for PD patients. When the disease gets detected premature identification, the patients can take proper medication and lead their life happily. However, the premature detection of PD is very challenging. The premature identification symptoms that can be detected for PD developing patients is sleep disorder. When the person gets sleep disorder, a high number of awakenings or less hours of sleep indicates the chances of PD symptoms, thus the sleep efficiency of the person is considered to be vital for detecting PD at early stages.

Parkinson’s disease can be diagnosed through a number of ways including speech, handwriting, gait and neuroimaging data, Electroencephalography (EEG) signals etc. These techniques can be detected once the symptoms of PD are detected by the persons, however, sleep disorder is one of the premature identification symptoms of having PD possibility, which is not addressed in the existing works. When detected late the severity of PD is high as it can cause motor impairment and make the curing process more difficult. Though the clinical study indicated the challenges in speech is the beginning stage of patients with PD, then it gradually affects the gait and motor impairments. Thus it is necessary to find the symptoms at premature identification to overcome the problem of gradual impairment losses for the patients with PD.

Simple components connected together to make a large complex system in real the world, which includes human brain, immune system, ant colonies, internet, economics and markets, this infers that the global behaviour making very

complex and makes real work simple component interactions. Identifying the network attributes like distributions, clustering helps in identifying more effective solutions, similarly, PD is the neurological disorder connecting multiple systems, the speech and gait movement can be considered as the complex system related to the neurological environment. PD is caused due to low dopamine level in the brain, this affects the complex systems like speech, gait, and movement. The severity caused to the PD cases may differ from one another depending on the affected levels of dopaminergic neuron death. Curing the PD is not possible, however if premature identification by proper medication and changing in the lifestyle reduces the severity of the disease and reduces the progress of the disease.

This work proposal includes

- The premature identification of Parkinson's Disease (PD) is proposed from sleep disorder among the people. The sleep efficiency gauges the good lifestyle practice. If the person is affected with sleep disorder or frequent awakening, the PD is prevalent.
- The prediction of PD is proposed with ensemble machine learning techniques considering the feature engineering to enhance the performance of the model.
- The proposed work contributes to premature identification of symptoms of PD through sleep efficiency analysis, thus the PD score data is computed from sleep efficiency
- This work extended for using three ML models Extra trees, AdaBoost and gradient Boosting models for PD prediction
- This work contributes to the high accuracy detection of disease using feature selection and hyper parameter tuning the ML models.
- The best fit parameters for the ML models are computed using Grid Search Cross Validation over five times and controlling the randomness of the train and test split the experiments are conducted and computed more accurate classification results.

When the PD progresses, the complications for the PD patients worsens and affects the normal life quality, the treatment of PD included dopamine drugs increases the level of neurotransmitter, which can exist for less period and these drugs do not change the neurodegenerative progress. Thus it is crucial to find alternate rehabilitation therapies, which can prevent the disease worsening. It is more challenging to identify the disease at earlier stages. In this work, proposed premature identification of patients with PD using sleep efficiency data.

This work addressed the problem of premature identification of PD from sleep efficiency data using ensemble machine learning models, the feature selection is performed for effective feature selection and hyper parameter tuning of machine learning models to get the best fit parameters for prediction.

The rest of the study is structured so that the literature review and pertinent research on Parkinson's disease prediction using artificial intelligence approaches are presented in Chapter 2. In Chapter 3, the suggested methods and optimization algorithms are discussed. Chapter 4 discusses the results of the PD prediction. Chapter 5 offers the investigation's conclusions as well as possible areas for development.

RELATED WORK

Detection of Parkinson's Disease (PD) is studied based on different data types including speech representations and phonation, speech tremor, electroencephalographic(EEG) data, wearable sensor like wearable armbands and a force-sensor, Gait analysis preformed. These works represent the detection of PD effectively, however, the premature identification of disease is still to be addressed. The sleeping disorder is one of the earlier symptoms of PD, which has to be explored in the analysis of PD.

Parkinson's disease detection using multimodal data is proposed in the work [2], which considered the speech signals and gait biomarkers for the analysis of PD. This work considering the multimodal data helped in identifying the PD severity, the classification is performed as three stages, Low, intermediate and severe. The input signals of speech and

Gait are pre-processed and generated visibility graphs, then feature level fusion is performed before applying ML models. There are eight ML models applied for classification of PD as three stages. Evaluation of results observed that Random forest classifier has outperformed the other models in accuracy is around 83.9%.

Electroencephalographic (EEG) recording is explored as the major source of PD detection, the work [3] proposed search algorithm, which selected the optimized set of channels for classification of PD. This work considered 60 channel EEG data collected from 20 control and 20 PD patients. The budget based greedy selection algorithm used to select the channel and classification. Evaluation of results showed that the AUC score of 0.71 is achieved through this proposed algorithm.

The work [4] proposed the ML model for differential diagnosis of Parkinson's patients and Essential tremor. The angular velocity signals are captured in rest and posture positions. This work addressed the Kinematic analysis, which classified the Healthy/Normal and Trembling cases and classified between patients with PD and Essential tremor. The extracted kinematic features include median power frequency, power bandwidth, peak power frequency, harmonic index etc. This work used various ML models for classification, the average accuracy achieved is 97.2% for healthy and trembling cases, average accuracy of 77.8% for PD and essential tremor case classifications.

Parkinson's disease detection from speech phonation and articulation is proposed in [5], this work used phonological features, glottal features and MFCC features from audio data. The temporal characteristics of attributes are gained through Convolutional Neural Network (CNN). The extracted features are undergone feature fusion process to ensure the performance of regression tasks at each segment level prediction. Experiment results showed the lesser RMSE computed for speech 11.7, followed by voice impairment 14.2, overall severity detected 11.8.

Premature detection of PD is proposed in the paper [6] exploiting Neural Network (NN) models, the dataset collected from Normal and PD cases, the severity of disease is detected as four stages. Dataset used is 6540 gait cycles collected from 48 subjects for this study, the lower arm band gait data is collected. The NN model has classified the PD and non-PD as well as stages of PD. Identifying the stages of PD helps in premature detection of PD. Evaluation of results observed that PD and non-PD detected at accuracy of 92.72% and stages of PD detected at accuracy of 99.67%

Another work [7] represented the premature identification of PD using machine learning and deep learning models using premotor features. This work considered the Cerebrospinal fluid, rapid eye movement and Parkinson's Progression Markers Initiative (PPMI) dataset for this study. Evaluation of the work showed that deep learning models have gained the maximum accuracy of 96.68% for classification of PD and non-PD over other machine learning models used. In the used ML models, Logistic regression was performing better and accuracy computed is 96%.

This literature survey for Parkinson's disease detection using AI techniques including different types of data like wearable sensor, gait representation, speech tremor are considered in these studies. It is observed that sleep disorder data is less explored for PD detection. Most of the studies identified the PD and non-PD subjects, however the premature identification of disease is less explored. Thus it is more challenging to identify the PD at premature identification stages. In this study, sleep efficiency data is used for identifying the disease at premature identification.

PROPOSED WORK

The proposed work is Parkinson's disease detection from the sleep efficiency of a person using ensemble machine learning techniques. The sleep disorder and frequent awakening are the premature identification symptoms of the PD, thus if analyzed, the disease can be detected earlier to prevent serious conditions as the disease progresses. This technique will provide a new lease of life to patients, as it accurately classifies PD using sleep data. This work addressed premature identification of PD prediction from sleep efficiency data using ML models Extreme Gradient Boosting, AdaBoost and Extra Trees models. The features are selected using wrapper based technique Recursive feature elimination and hyper parameter tuning of the ML model is performed for better performance of the model. The proposed work aimed at accurate detection of PD from sleep efficiency data, this work proposed to overcome the problem and challenges of premature identification.

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

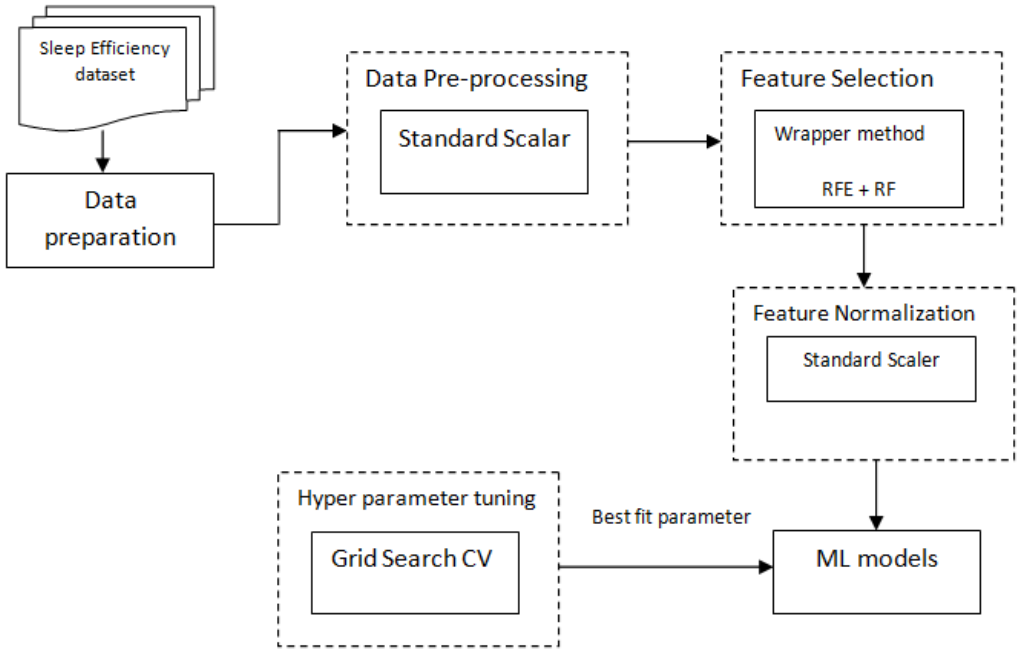

Figure 1: System architecture of premature identification of Parkinson's Disease

DATASET DETAILS

Details on a group of people's sleeping habits are included in the dataset. A distinct "Subject ID" is assigned to each instance in the data, together with the data's age and gender. Each person's sleep length is determined by their "Bedtime," "Wakeup time," and "Sleep duration." The amount of time spent sleeping is indicated by the "Sleep efficiency" score. The stages of sleep are mentioned in the "REM sleep percentage," "Deep sleep percentage," and "Light sleep percentage" charts. The "Awakenings" are just the amount of times the individual wakes up while they are asleep. The dataset also includes the person's smoking status, frequency of exercise, and amount of alcohol and caffeine consumed in the 24 hours before bedtime. There are a total of 14 features and 452 instances available in the dataset.

PD SCORE COMPUTATION

Using the weighted sum of selected features, created a score or probability metric that classifies PD vs. Non-PD. The formula considered to arrive the PD score is given below

$$PD\_Score =$$

$$0.4\times(1-SleepEfficiency)+0.3\times Awakenings+0.2\times(100-REMSleep\ \%)+0.1\times(100-Deep\ Sleep\ \%)$$

- Sleep Efficiency: Inverts efficiency so higher values align with PD likelihood.
- Awakenings: More awakenings indicate more disrupted sleep, likely in PD.
- 100−REM Sleep Percentage: Lower REM percentage indicates a likelihood of PD.
- 100−Deep Sleep Percentage: Lower deep sleep percentage is more common in PD.

By setting a threshold on the PD_Score, individuals are labelled as likely PD or Non-PD. This is an assumption-based approach, done to label the dataset as PD and non-PD cased.

Visualization of PD data

The visualization of on Parkinson's disease dataset considering sleep efficiency is performed. The below Figure 1 shows the frequency of PD score. The PD score computed for the dataset exists in the range from 18 to 26. It is observed from Figure 1 that the number of cases with PD score 20 is around 140 and the second largest frequency is observed in 19,

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

which is around 90 persons. The third distribution seen around PD score 21 is 78 persons. The mean of the distribution observed is 20.7 as shown in the figure. There are 55 instances available at PD score around 23.5. The high the PD score, the prevalence of disease is high. The low the PD score the persons instances are considered to be non-PD.

Figure 2 represents the distribution of frequent awakening in the dataset, they are given as categorical range from 0 to 4, where 0 number of awakenings to 4 number of awakenings. The frequent awakenings the PD prevalence is high, whereas less the awakening is non-PD. In this dataset, the frequent awakening 1 has a total frequent PD cases 28 and non-PD cases 122. The frequent awakenings 4 has the PD cases 40 and non-PD cases 22. This clearly represents that higher the frequent awakenings has the possibility of having PD and other factors are also considered in the dataset like sleep efficiency to arrive at the PD score and infer the PD availability.

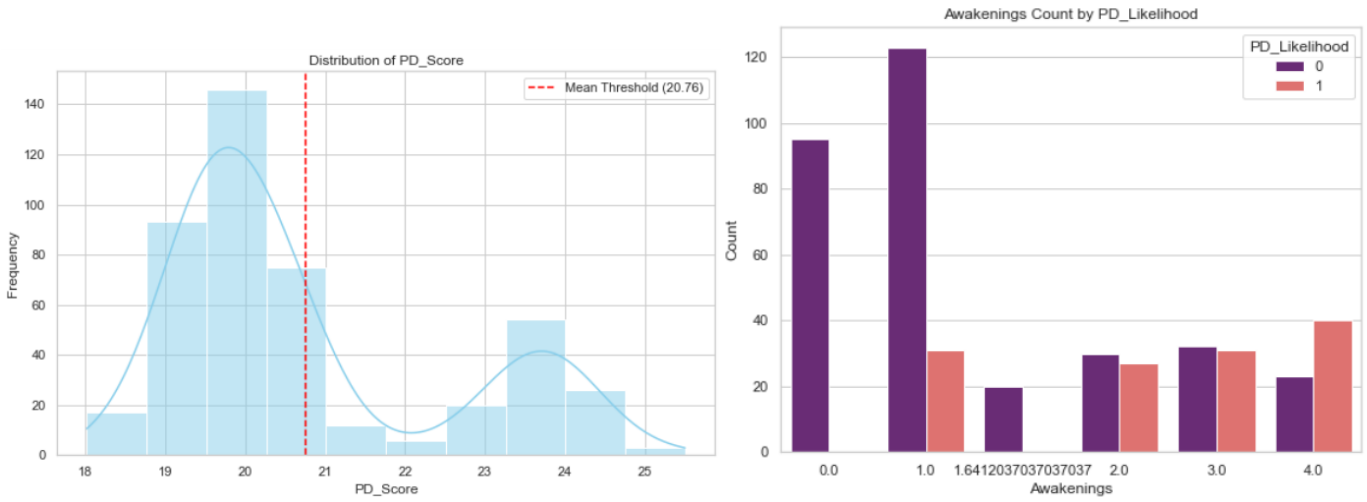

Figure 2: (a)Frequency of PD Score (b) Distribution of Frequent Awakenings in dataset

METHODOLOGY

The proposed methodology is early prediction of Parkinson's disease from sleep efficiency dataset as binary classification of PD and non-PD cases. Ensemble ML technique were exploited in this study with effective features selected using wrapper based feature selection technique 'Recursive Feature Elimination' (RFE) and the ensemble ML models are hyper tuned using Grid Search Cross Validation (GSCV) technique. The data is normalized in the pre-processing to give normalized values to the ML model. Figure 1 shows the overall system architecture of the proposed PD classification using ensemble models. In this work, the tree based model and boosting models were used for classification. The tree model used is extra trees, boosting models used are Adaboost and Extreme gradient boosting models. These models are experiments after performing hyper parameter tuning to get the better improved model.

Data Normalization

Sleep efficiency data has few attributes namely REM sleep percentage, Deep sleep percentage, and Light sleep percentage as continuous values with large range, thus normalization is applied to the dataset. The data is normalized using z-score computation, which is by applying StandardScaler() to fit and transform to normalized data so that each attribute in sleep efficiency data has mean 0 and standard deviation 1.

$$z = \frac{x - \mu}{\sigma} \quad \text{-(1)}$$

In the above formula, where x is the given data point, μ is the mean value and σ is standard deviation.

FEATURE SELECTION

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

As a part of data pre-processing, feature selection is performed on sleep efficiency data using 'Recursive Feature Elimination (RFE)', the number of features selected is 10 based on the ranking score arrived by the technique. The features selected for the model include age, sleep duration, sleep efficiency, REM sleep %, deep sleep %, light sleep %,awakening, alcohol, exercise frequency and PD_Score.

Figure 3 shows the feature selection process overview of RFE technique, Random Forest model is considered as an estimator for RFE technique, which generates a ranking to select the best 10 features from the given dataset.

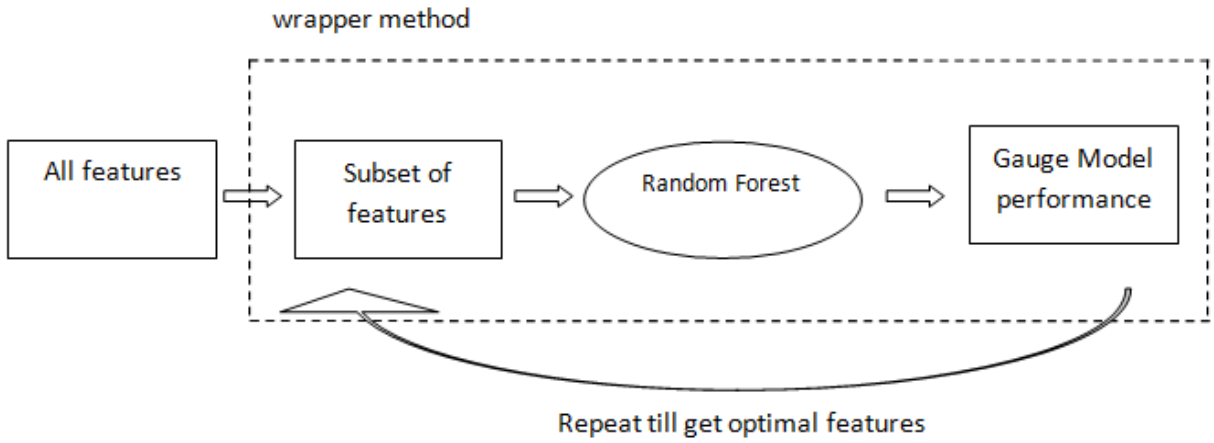

Figure 3: RFE estimator Overview

MODEL OPTIMIZATION

The ensemble ML models extra trees, ADB and XGB are hyper parameters optimized through Grid search cross validation. In the extra trees model 6 parameters are taken for optimizing including number of estimator, max depth of tree, minimum sample split and leaf, max features, the cross validation is performed 5 times to get best fit parameters. For the XGB model, three parameters are tuned including number of estimators, max depth, and gamma values. Similarly for ADB three parameters tuned are number of estimators, learning rate, and best estimators.

Table 1 shows the hyper parameter tuned for ensemble models using Grid Search Cross Validation (GSCV) technique. Cross validation of 5 folds is used.

Table 1: Hyper parameter search space for PD prediction

Algorithm	Parameters and Value for optimization
Extra Trees	n_estimators:Randint(100,1000) max_depth:Randint(5, 30) min_samples_split:randint(2, 20) min_samples_leaf:randint(1, 20) Bootstrap:[True, False] max_features:['auto', 'sqrt', 'log2']
XGB	n_estimators:[100, 200,300,400,500] max_depth:[3, 5, 7] Gamma:[0, 0.1, 0.2]
ADB	n_estimators:[50, 100] learning_rate:[2.0, 2.5] base_estimator__max_depth:[1, 2, 3]

RESULTS AND DISCUSSIONS

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The experiment is conducted on ensemble ML models, Extra Gradient Boosting, AdaBoost and Extreme Gradient Boosting for classification of sleep efficiency data as PD and non-PD based on PD score computed. The experimental set up is performed with 80% training and 20% test dataset.

Table 2: Comparison of proposed models for Chronic Kidney disease prediction

Algorithm	Accuracy (%)
Extra Trees+ RFE+HT	96.70
XGB + RFE +HT	99.11
AdaBoost Model+ RFE+HT	98.74

The below table shows results of ML models, HT- represents Hyper tuning. The accuracy of the AdaBoost Model is 99.11% ,the XGBoosting model is 98.74% and the Extra tree model is 96.7%.

Table 3: Performance comparison of proposed models for Chronic Kidney disease prediction

Algorithm	MAE
Extra Trees+ RFE+HT	0.0329
XGB + RFE +HT	0.0088
AdaBoost Model+ RFE+HT	0.0125

The mean absolute error is computed for the model, Evaluation of results observed that AdaBoost model showed less error around 0.0088. For the XGBoost model MAE is 0.0125 and Extra Trees model is 0.0329 respectively.

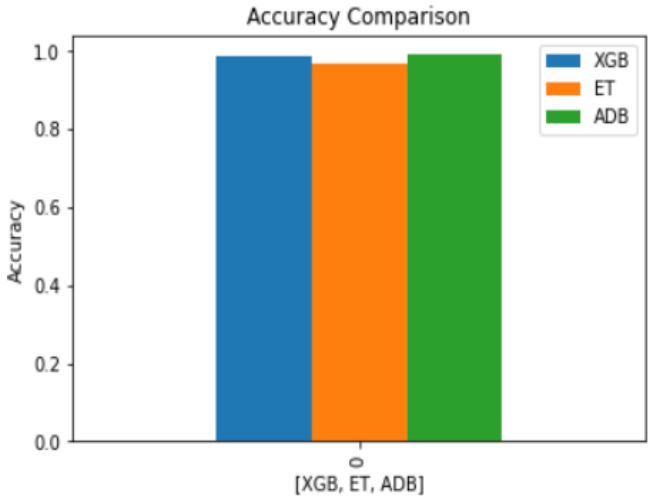

Figure 3: Accuracy of proposed ML models for Parkinson's Disease classification

Figure 3 represents the performance comparison of different ensemble learning for PD premature detection. From the figure, it is observed that the XBG model showed higher classification accuracy of 99.11% than other two ML models.

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

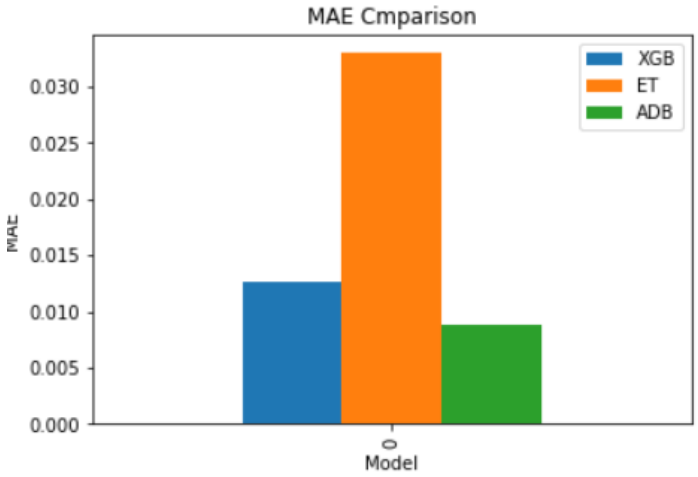

Figure 3: MAE of proposed ML models for Parkinson's Disease classification

Figure 4 represents the error metric MAE computed for different ensemble mechanism for PD premature detection. From the figure, it is evident that the XBG model has a lower MAE value of around 0.0088.

CONCLUSIONS

In this work, Parkinson's disease prediction from sleep efficiency is proposed using machine learning. The feature selection is performed using Recursive Feature Elimination (RFE) and the model is hyper tuned using Grid Search Cross Validation (GSCV). The model is optimized to get the best performance from the ML models. The proposed work showed more accurate classification using the proposed novel ML models. The sleep efficiency data helps to identify the persons with Parkinson's at early stages so that they can take proper medications. Experimental results showed that the model has achieved high accuracy of 99.11% for binary classification of Parkinson's disease.

As a further enhancement, this work can be extended to analyze the features using other techniques including Genetic Algorithm and PCA analysis. Also this work can be further extended to use Deep learning models.

REFERENCES

[1] https://www.who.int/news-room/fact-sheets/detail/parkinson-disease
[2] K. D. Raj, G. J. Lal, E. A. Gopalakrishnan, V. Sowmya and J. R. Orozco-Arroyave, "A Visibility Graph Approach for Multi-Stage Classification of Parkinson's Disease Using Multimodal Data," in IEEE Access, vol. 12, pp. 87077-87096, 2024, doi: 10.1109/ACCESS.2024.3416444.
[3] I. Suuronen, A. Airola, T. Pahikkala, M. Murtojärvi, V. Kaasinen and H. Railo, "Budget-Based Classification of Parkinson's Disease From Resting State EEG," in IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 8, pp. 3740-3747, Aug. 2023, doi: 10.1109/JBHI.2023.3235040.
[4] J. D. L. Duque, A. J. S. Egea, T. Reeb, H. A. G. Rojas and A. M. González-Vargas, "Angular Velocity Analysis Boosted by Machine Learning for Helping in the Differential Diagnosis of Parkinson's Disease and Essential Tremor," in IEEE Access, vol. 8, pp. 88866-88875, 2020, doi: 10.1109/ACCESS.2020.2993647.
[5] Y. Liu, M. K. Reddy, N. Penttilä, T. Ihalainen, P. Alku and O. Räsänen, "Automatic Assessment of Parkinson's Disease Using Speech Representations of Phonation and Articulation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 242-255, 2023, doi: 10.1109/TASLP.2022.3212829.
[6] C. -H. Lin, F. -C. Wang, T. -Y. Kuo, P. -W. Huang, S. -F. Chen and L. -C. Fu, "Early Detection of Parkinson's Disease by Neural Network Models," in IEEE Access, vol. 10, pp. 19033-19044, 2022, doi: 10.1109/ACCESS.2022.3150774.
[7] W. Wang, J. Lee, F. Harrou and Y. Sun, "Early Detection of Parkinson's Disease Using Deep Learning and Machine Learning," in IEEE Access, vol. 8, pp. 147635-147646, 2020, doi: 10.1109/ACCESS.2020.3016062.
[8] H. Gunduz, "Deep Learning-Based Parkinson's Disease Classification Using Vocal Feature Sets," in IEEE Access, vol. 7, pp. 115540-115551, 2019, doi: 10.1109/ACCESS.2019.2936564.

Research Article

[9] R. LeMoyne, C. Coroian and T. Mastroianni, "Quantification of Parkinson's disease characteristics using wireless accelerometers," 2009 ICME International Conference on Complex Medical Engineering, Tempe, AZ, USA, 2009, pp. 1-5, doi: 10.1109/ICCME.2009.4906657.

[10] R. LeMoyne, T. Mastroianni, M. Cozza, C. Coroian and W. Grundfest, "Implementation of an iPhone for characterizing Parkinson's disease tremor through a wireless accelerometer application," 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, Buenos Aires, Argentina, 2010, pp. 4954-4958, doi: 10.1109/IEMBS.2010.5627240.

[11] F. M. Khan, M. Barnathan, M. Montgomery, S. Myers, L. Côté and S. Loftus, "A Wearable Accelerometer System for Unobtrusive Monitoring of Parkinson's Diease Motor Symptoms," 2014 IEEE International Conference on Bioinformatics and Bioengineering, Boca Raton, FL, USA, 2014, pp. 120-125, doi: 10.1109/BIBE.2014.18.

[12] Camacho M, Wilms M, Mouches P, Almgren H, Souza R, Camicioli R, Ismail Z, Monchi O, Forkert ND. Explainable classification of Parkinson's disease using deep learning trained on a large multi-center database of T1-weighted MRI datasets. Neuroimage Clin. 2023;38:103405. doi: 10.1016/j.nicl.2023.103405. Epub 2023 Apr 17. PMID: 37079936; PMCID: PMC10148079.