Original Article

Early Warning Signs Of Parkinson's Disease Prediction Using Machine Learning Technique

Pawan Kumar Mall[1] , Rajesh Kumar Yadav[2] , Arun Kumar Rai[3] , Vipul Narayan[4] , Swapnita Srivastava[5]

[1]Lovely Professional University, Jalandhar- Delhi G.T. Road, Phagwara, Punjab 144411 pawankumar.mall@gmail.com

[2]Directorate of Education, GNCT of Delhi Old Secretariat, Delhi 54 rajeshyadav491@gmail.com

[3]Greater Noida Institute of Technology, Knowledge Park II, Greater Noida, Uttar Pradesh, India-201310 arun.akrai@gmail.com

[4]Galgotia University Gautam buddh Nagar, Uttar Pradesh vipulupsainian2470@gmail.com

[5]Galgotia University Gautam buddh Nagar, Uttar Pradesh swapnitasrivastava@gmail.com

DOI: 10.47750/pnr.2022.13.S10.579

Abstract

Brain cells breakdown is primary cause of Parkinson's disease (PD) that create dopamine. A neurotransmitter that allows brain cells to connect with each other. Dopamine-producing cells in the brain are in control of adaptability, movement regulation, and fluency. Active ageing is a concept that has arisen to optimise many areas of fitness prospects throughout the ageing development in order to improve eminence of life. However, best initiatives have focused on normal ageing, with little attention dedicated to the elderly (ageing) suffering from a chronic condition such as PD. These tactics may be able to assist people with PD in better managing their illness in the context of active ageing. The datasets have extensively depicted parkinson disease by utilising a range of data-derived perspectives. According to the results, our suggested technique beats other techniques. In order to achive our aim we have suggested an ensemble learning technique. The suggested model outperforms existing machine learning approaches such as SVM(support vector machine), KNN(K-nearest neighbour), RF(Random-Forest), DT(Decision-Tree), MLP(multilayer perceptron), (SC)StackingClassifier, (LR)Logistic-Regression. when accuracy, matthews correlation coefficient (MCC), and f1score are calculated. According to the results of our research, the technique we described, the ensemble model, outperforms other machine learning models. We achieved 94.87% accuracy, 81.99% MCC , and 94.52% f1score.

1. **Introduction:** The World Health Organization defined "active ageing" in the late 1990s as "the process of optimising possibilities for health, involvement, and security in order to improve life style of ageing people". The most of which have motivated solely on the general people, with little consideration dedicated to the elderly suffering from a chronic condition such as Parkinson's disease (PD). In most communities, genetic reasons connected to recognised PD genes explain 3-5% of the heritable risk of disease, whereas 90 genetic risk variations explain 16-36% of the heritable risk of non-monogenic Parkinson's disease. Having a family with PD, non smoking, and constipation all increase the chance of PD by at least double [5]. PD is recurrent movement ou scondition worldwide, affecting around 1% of persons over the age of 60 [21]. At the moment, no therapy can halt or stop the course of PD, although numerous intriguing techniques are being investigated for disease-modifying potential in light of recent insights into genetic origins and processes of neuronal death [1].

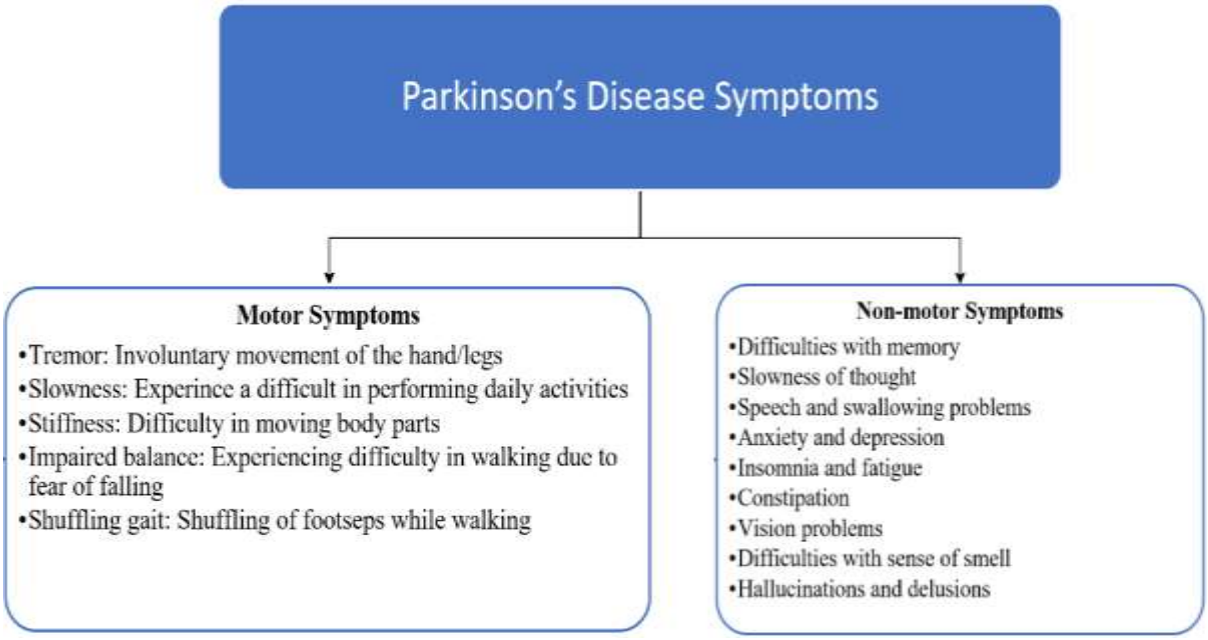

Figure 1: Symtoms of Parkinson disease

The main contributions of this work can be encapsulated as follow:

- PD-Detector solution capable of efficiently discriminating distinct PD with high accuracy;
- PD-Detector model development has tremendous potential for use in hospital applications;

The outline of this research work as follows:

- At first, we have presented the literature survey
- Next is the detail description of dataset
- Next we have introduced our proposed approach.
- The next section provides detailed experimental results.
- Finally, Sections five conclude this work and discussed the future direction of this research work.

2. **Related Work:** In [1], auhors have implemented Wavelet analysis support vector machine (SVM) as paired approach for efficient classification accuracy of 90.32%. In [2], author have investigated several machine learning models random forests(RF), logistic regression(LR), XGboost and support vector machines(SVM). Model-free machine learning model achived 80% accuracy. In [20], This pilot study's goal was to apply machine learning (ML) classification to evaluate the merits and applicability of neurocognitive variables in relation to PD and its stages using tablet-based evaluations and self-reported metrics. With an accuracy of 92.6%, decision tree classification of sensor-based information allowed for the separation of PD from healthy controls. In [3], The deep CNN framework with pre - trained models VGG16 and layout beat other methods because it successfully differentiated between patients with PD and normal subjects, caught important imaging information, and had the best classification accuracy. In [17], The effectiveness of proposed Ensemble learning classifiersis compare with RF, and Boosting DT, LR, and KNN. In [18], authors have trained RF and SVM models with normalized data and evaluated using cross validation and AUC. In general SVM performed better than RF. In [16], authors have implemented different types of MLA, such as SVM, LR, Discriminant Analysis (DA), KNN, DT, RF, Bagging tree (BT), Naïve Bayes (NB), and AdaBoost. These ML algorithms were implemented using Parkinson's dataset with 44 different features and 240 speech measurements. The accuracy performance evaluation parameters is use to validate and it is conclude that the KNN model yielded the maximum accuracy 97.22% and F1-score value of 97.30%.

3. Proposed Model:

The suggested technique, which includes data gathering, preprocessing, machine learning classifiers, and performance indicators, is discussed in this section. Vote count model (VCM) approach has been utilised to classify these issues, and data pre-processing has been employed to put datasets in the proper manner to enhance classification work. In this study, a novel VCM approach is presented and evaluated in a challenging classification PD data set where the majority of classifiers perform averagely and achieve less than 70% accuracy. The dataset is related to PD, a respiratory infection problem with a significant mortality risk. Figure 2. Depicts the complete architecture of proposed model.

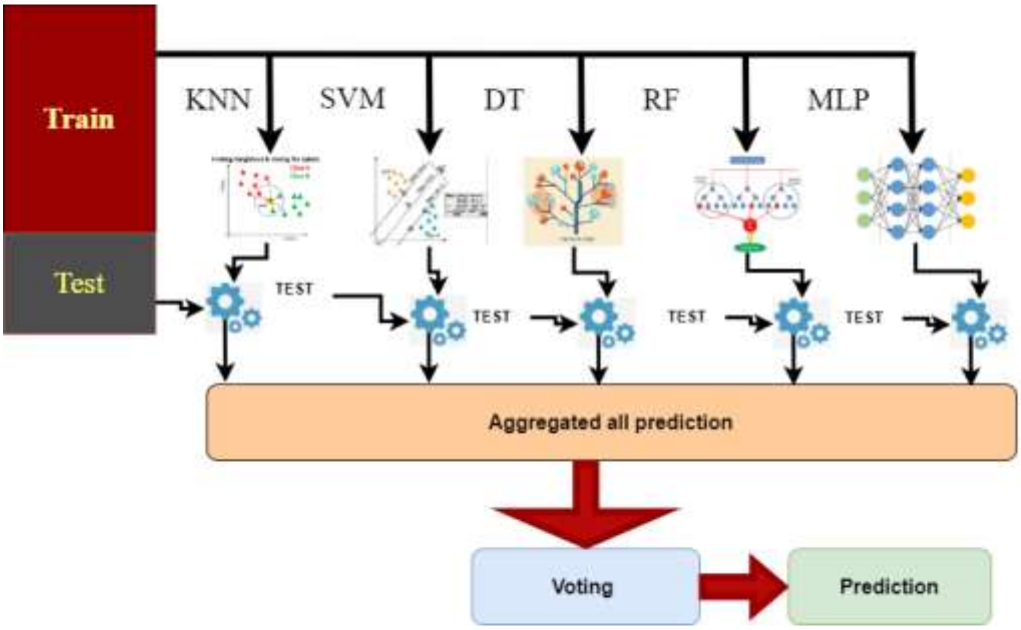

Figure 2. Proposed Architecture

3.1 Data acquisition and pre-processing:

The University of California, Irvine's (UCI) Machine Learning Repository provided the dataset [6]. Data Set for PD. The information gathered on patients in the actual world is frequently inconsistent, lacking, and likely to include several inaccuracies. Raw data is transformed using data pre-processing techniques into a machine-readable format. Data pre-processing methods can address these concerns. Age is the only characteristic for which the MinMax Scaler approach is used because all other features are in single digits. Therefore, the scale age feature within the raging is necessary. Figure 3. provides PD dataset details.

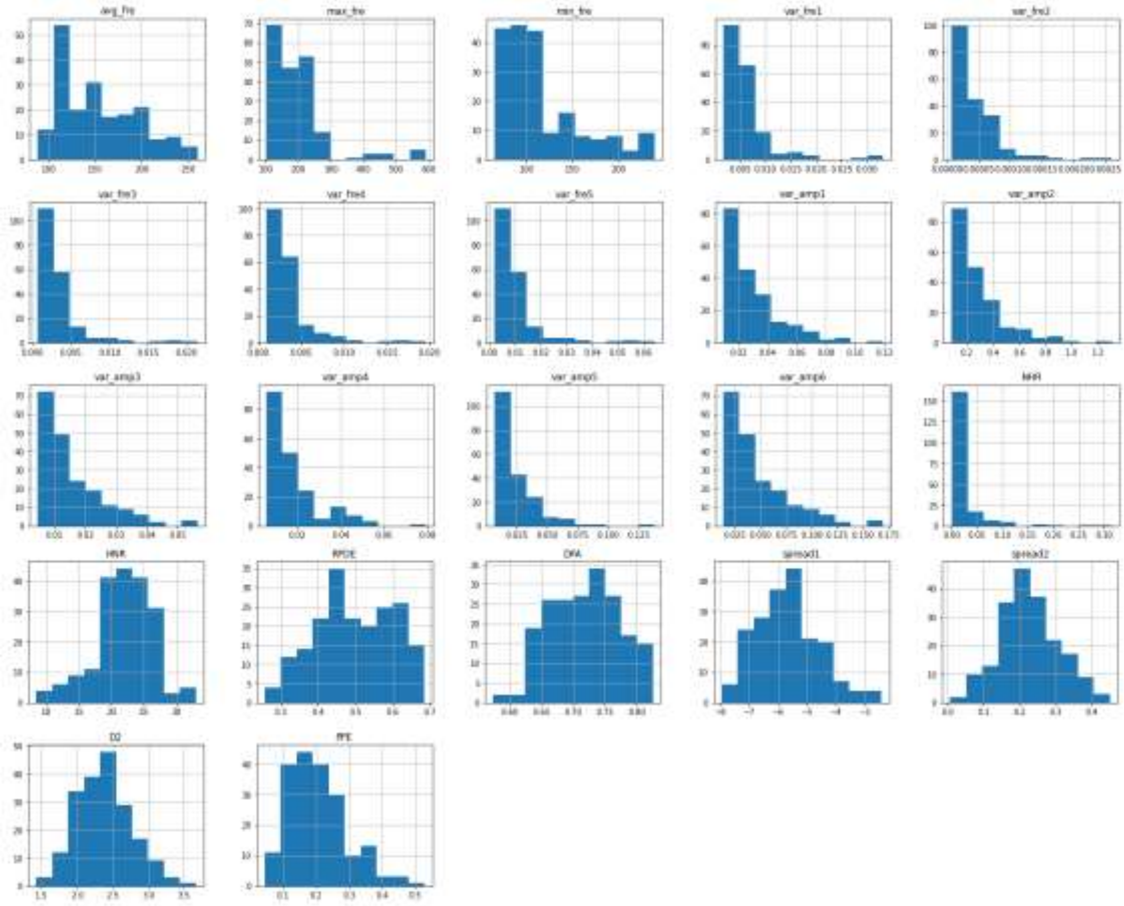

Figure 3. PD dataset details:

3.2 Decision Tree (DT):

A broad predictive modelling method called decision tree analysis has a wide range of applications. Decision trees are often built using an algorithmic method that finds ways to divide a data set depending on several criteria. It is among the most popular and useful techniques for supervised learning. A non-parametric supervised learning technique called decision trees is utilised for both classification and regression applications. The objective is to learn straightforward decision rules derived from the data characteristics in order to build a model that predicts the value of a target variable.[19].

3.3 KNN (K- Nearest Neighbors) Algorithm:

Since KNN depends on computing distances, it is quite sensitive to the quantity of training data. This method does not retrieve results based on presumptions regarding the training set of data. The KNN method is regarded as being quite good at addressing issues that come with non-linear data points, even if this may not be the overall situation when you examine other supervised learning algorithms [10] [12].

3.4 SVM (Support Vector Machine) Algorithm:

In SVM technique the data is classified by plotting the raw data as dots into an n-dimensional (where n represent number of features). Now the data can be easily classified as each feature's value is connected to a specific coordinate .The data may be grouped and displayed on a graph using classifiers [14] .

3.5. Random Forest:

The random forest classifier comes under supervised learning method and is used for classification and regression issues. Due to its great flexibility and simplicity of use, it is one of the most preferred machine learning algorithms. Compared to the majority of non-linear classifiers, the random forest approach is substantially more accurate. Because it generates its output using many decision trees, this method is also quite reliable. Because the random forest classifier averages all predictions, eliminating out biases and resolving the overfitting issue, it does not experience the overfitting problem [8] [11].

3.6 MLP:

The much more complicated design of artificial neural networks is defined by the multi-layer perceptron. It is largely constructed from several perceptron layers. A directed graph connecting the input and output layers of an MLP is made up of many layers of input nodes. Backpropagation is used by MLP to train the network. A deep learning technique is MLP [13] [9].

3.7 Proposed Approach: Our proposed model is based on ensemble technique. When numerous distinct model types are fitted to the same data, stacking is used to understand how to integrate the predictions most effectively. Algorithm 1 provides the procedure for the PD prediction on ensemble technique [22][23][24].

Algorithm 1: Procedure for PD prediction based on ensemble technique:

```
INPUT:TestData, N, M
OUTPUT: ODL
Initialization;
TestData= testing dataset,
N = Number of Standard Machine learning model,
M = Size of test dataset,
ODL = predicted output,
M_size ← size of (testdata)
Begin
for k=1 to M_size do
        For i=1 to N do
            ODL[k][i] = class lable of predicted model
            If ODL[k][i] is Class1
            Vote_c1= Vote_c1+1
            Else
            Vote_c2= Vote_c2+1
         End for
          IF Vote_c1> Vote_c2
          ODL[k] = Class1
          ELSE IF Vote_c1= Vote_c2
          ODL[k] = Class1
          ELSE
          ODL[k] = Class2
End for
End
```

4. Result:

In this section we are going to analysis accuracy of the hybrid model's classification is the primary focus of this research paper's primary purpose, which is to increase that model's classification accuracy. The experiment is carried out in virtual machine configure with Ubuntu and python 3.6 install. The experimental assessment of the hybrid architecture that was put into place was carried out in terms of the following parameters. Figure 4. Shows correlation matrix to identify the highly coreleted features.

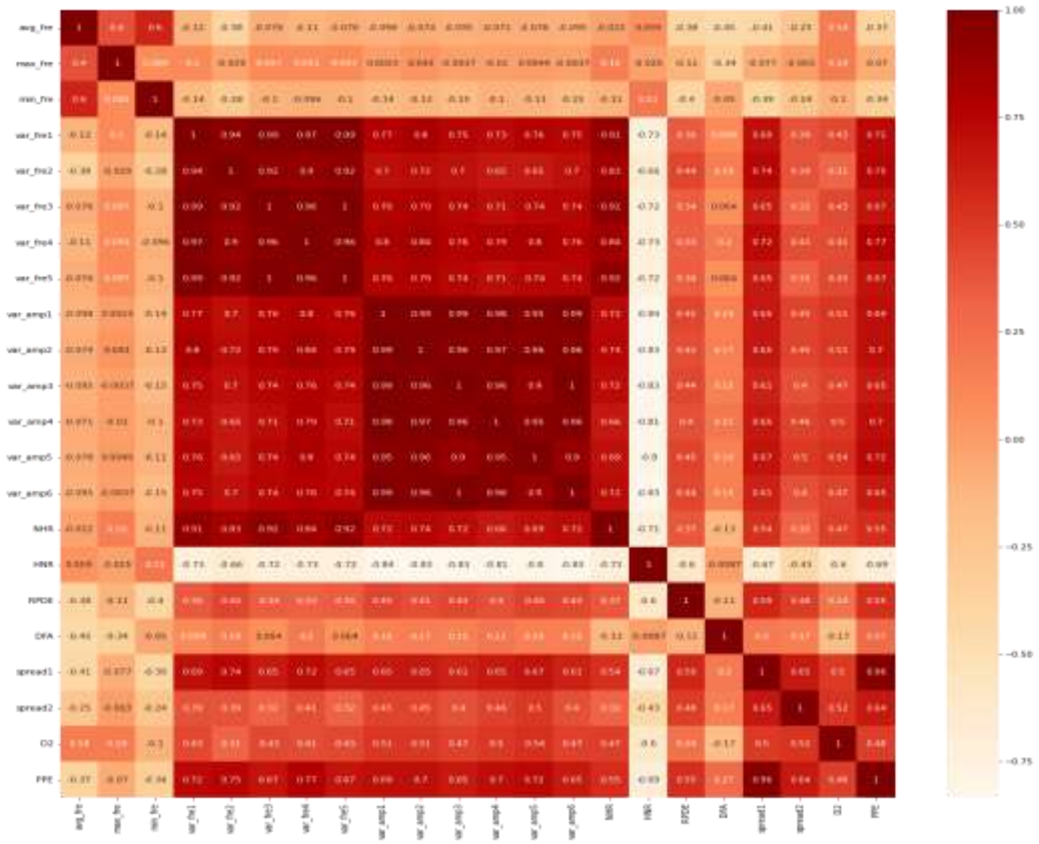

Figure 4. correlation matrix

We evaluate accuracy, MCC, and F1score, and offer the comparative findings for a variety of assaults since intrusion detection systems need a very high detection rate and a low false alarm rate.

- Accuracy: The Accuracy [6] refers to how close a measurement is to its true value. The 'Accuracy' is evaluated as depicted in equation (1) as follows:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

- F1 Score (F1): The highest F1[4] indicates flawless recall and precision, whereas the lowest F1 value suggests no recall or precision. The 'F1' is evaluated as depicted in equation (2) as follows:

$$F1 = \frac{2TP}{2TP+FP+FN} \qquad (2)$$

- **Matthews correlation coefficient (MCC)** [7]: This statistic is especially helpful when there is an imbalance between the two classes, meaning that one class appears substantially more frequently than the other. The range of the MCC value is -1 to 1, with 1 denoting entire disagreement between predicted classes and actual classes, 0 denoting completely random guessing, and 1 denoting perfect agreement between anticipated classes and actual classes.

$$\mathbf{MCC} = (TP*TN–FP*FN)/(TP+FP)(TP+FN)(TN+FP)(TN+FN) \qquad (3)$$

$$\text{where TP is true positive, TN is true negative, FP isfalse positive,}$$

$$\text{FN is false negative}$$

Table 1 provides the comparative analysis of ML Model on train data and table 2 provides comparative analysis of ML Model on train data.

Table 1. Comparative analysis of ML Model on train data

Models	**Accuracy**	**MCC**	**F1score**
KNN	96.15	90.27	96.12

SVM	98.71	96.69	98.70
Decision Tree	95.51	88.79	95.56
Random forest	100	100	100
MLP	89.10	70.77	88.75
Proposed model	98.71	96.69	98.70

Table 2. Comparative analysis of ML Model on test data

Models	**Accuracy**	**MCC**	**F1score**
KNN	94.87	81.99	94.52
SVM	89.74	61.72	87.99
Decision Tree	84.61	59.77	85.81
Random forest	87.17	49.70	84.08
MLP	92.30	72.28	91.43
Nishat et.al [15]	93.39%	NA	NA
Proposed model	94.87	81.99	94.52

The Figure. 5. provides the comparative findings for a variety of assaults accuracy, MCC, and f1score, since PD prediction need a high accuracy, precision.

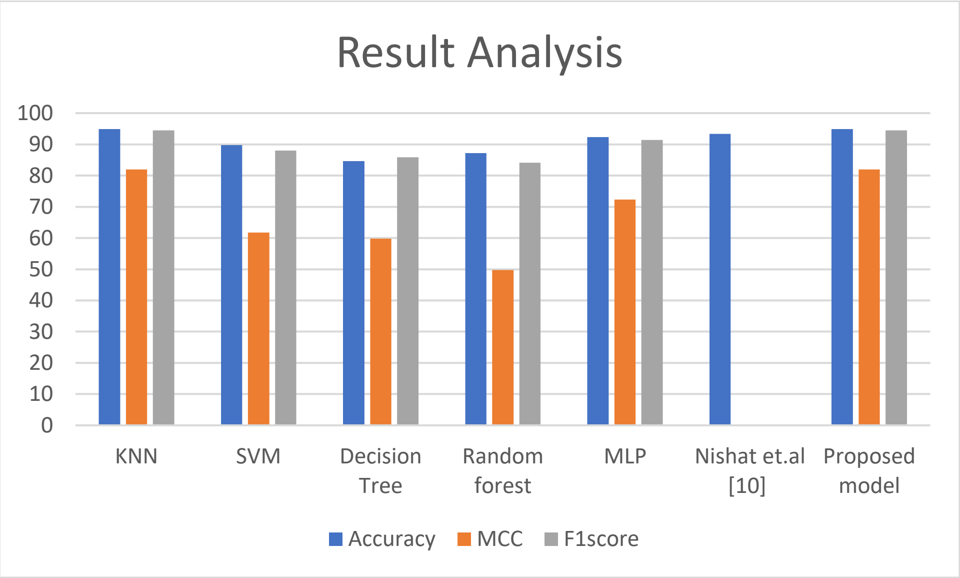

Figure 5. Result analysis of ML Model

According to the findings of our investigation, the approach that we have presented, ensemble model, performs better than other machine learning models. We were able to achieved 94.87% accuracy, 81.99% MCC , and 94.52% f1score. In all performance criteria, our proposed model outperformed the model proposed by [15]. As a result, the given model is trustworthy and suitable for quick and accurate PD patient screening. When training and testing our models, time complexity and structure complexity both increase significantly.

6. Conclusion:

Given that cardiac arrest caused almost 6, 17,000 fatalities in 2017, it is imperative that it receives special attention in today's society. To stop fatalities from occurring sooner, early PD prediction and protective actions are required. The outcomes show that our suggested method performs better than existing ML algorithms. The suggested model outperforms existing machine learning methods, including Decision Tree, SVM, KNN, Random Forest, and XG-Boost. Their accuracy, MCC, and f1score are measured as part of a comparative study. Our investigation revealed that our suggested approach performs better than other machine learning methods. According to the results of our analysis, the ensemble model technique that we have described outperforms alternative machine learning models. Our results showed We achieved a 94.87% accuracy, 81.99% MCC , and 94.52% f1score.

References:

1. Bloem, B.R. et al.: Parkinson's disease. Lancet. 397, 10291, 2284–2303 (2021). https://doi.org/https://doi.org/10.1016/S0140-6736(21)00218-X.
2. Gao, C. et al.: Model-based and Model-free Machine Learning Techniques for Diagnostic Prediction and Classification of Clinical Outcomes in Parkinson's Disease. Sci. Rep. 8, 1, 7129 (2018). https://doi.org/10.1038/s41598-018-24783-4.
3. Huang, G.-H. et al.: Multiclass machine learning classification of functional brain images for Parkinson's disease stage prediction. Stat. Anal. Data Min. ASA Data Sci. J. 13, 5, 508–523 (2020).
4. Irfan, D. et al.: Prediction of Quality Food Sale in Mart Using the AI-Based TOR Method. J. Food Qual. 2022, (2022).
5. Karapinar Senturk, Z.: Early diagnosis of Parkinson's disease using machine learning algorithms. Med. Hypotheses. 138, 109603 (2020). https://doi.org/https://doi.org/10.1016/j.mehy.2020.109603.
6. Mall, P.K. et al.: GLCM based feature extraction and medical X-RAY image classification using machine learning techniques. In: 2019 IEEE Conference on Information and Communication Technology. pp. 1–6 (2019).
7. Mall, P.K., Singh, P.K.: Explainable Deep Learning approach for Shoulder Abnormality Detection in X-Rays Dataset. Int. J. Next-Generation Comput. 13, 3, (2022).
8. Narayan, V. et al.: Energy efficient two tier cluster based protocol for wireless sensor network. In: 2020 international conference on electrical and electronics engineering (ICE3). pp. 574–579 (2020).
9. Narayan, V., Daniel, A.K.: A novel approach for cluster head selection using trust function in WSN. Scalable Comput. Pract. Exp. 22, 1, 1–13 (2021).
10. Narayan, V., Daniel, A.K.: CHHP: coverage optimization and hole healing protocol using sleep and wake-up concept for wireless sensor network. Int. J. Syst. Assur. Eng. Manag. 1–11 (2022).
11. Narayan, V., Daniel, A.K.: Design Consideration and Issues in Wireless Sensor Network Deployment.". Invertis J. Sci. \& Technol. 101 (2020).
12. Narayan, V., Daniel, A.K.: IOT Based Sensor Monitoring System for Smart Complex and Shopping Malls. In: International Conference on Mobile Networks and Management. pp. 344–354 (2021).
13. Narayan, V., Daniel, A.K.: Multi-Tier Cluster Based Smart Farming Using Wireless Sensor Network. In: 2020 5th International Conference on Computing, Communication and Security (ICCCS). pp. 1–5 (2020). https://doi.org/10.1109/ICCCS49678.2020.9277072.
14. Narayan, V., Daniel, A.K.: RBCHS: Region-based cluster head selection protocol in wireless sensor network. In: Proceedings of Integrated Intelligence Enable Networks and Computing. pp. 863–869 Springer (2021).
15. Nishat, M.M. et al.: Detection of Parkinson's Disease by Employing Boosting Algorithms. In: 2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR). pp. 1–7 (2021). https://doi.org/10.1109/ICIEVicIVPR52578.2021.9564108.
16. Ouhmida, A. et al.: Parkinson's disease classification using machine learning algorithms: performance analysis and comparison. In:

2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET). pp. 1–6 (2022). https://doi.org/10.1109/IRASET52964.2022.9738264.

17. Patra, A.K. et al.: Prediction of Parkinson's disease using Ensemble Machine Learning classification from acoustic analysis. J. Phys. Conf. Ser. 1372, 1, 12041 (2019). https://doi.org/10.1088/1742-6596/1372/1/012041.

18. Rehman, R.Z. et al.: Comparison of Walking Protocols and Gait Assessment Systems for Machine Learning-Based Classification of Parkinson's Disease, (2019). https://doi.org/10.3390/s19245363.

19. Song, Y.-Y., Ying, L.U.: Decision tree methods: applications for classification and prediction. Shanghai Arch. psychiatry. 27, 2, 130 (2015).

20. Templeton, J.M. et al.: Classification of Parkinson's disease and its stages using machine learning. Sci. Rep. 12, 1, 14036 (2022). https://doi.org/10.1038/s41598-022-18015-z.

21. Uwishema, O. et al.: The understanding of Parkinson's disease through genetics and new therapies. Brain Behav. 12, 5, e2577 (2022).

22. Narayan, Vipul, and A. K. Daniel. "CHOP: Maximum coverage optimization and resolve hole healing problem using sleep and wake-up technique for WSN." ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal 11.2 (2022): 159-178.

23. Narayan, Vipul, and A. K. Daniel. "Energy Efficient Protocol for Lifetime Prediction of Wireless Sensor Network using Multivariate Polynomial Regression Model." Journal of Scientific & Industrial Research 81.12 (2022): 1297-1309.

24. Narayan, Vipul, and A. K. Daniel. "FBCHS: Fuzzy Based Cluster Head Selection Protocol to Enhance Network Lifetime of WSN." ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal 11.3 (2022): 285-307.