

Using Machine Learning Algorithms to Predict Parkinson's Disease : A Comparative Analysis Based on Logistic Regression, Random Forest and XGBoost

Jinqiu Zhang *

Maynooth International Engineering College, Fuzhou University, Fuzhou, 350100, China

* Corresponding Author Email: 832203112@fzu.edu.cn

Abstract. Parkinson's disease, as a neurodegenerative disease, seriously affects the quality of life of patients. Early and accurate diagnosis is crucial for treatment. For comparison and prediction, this study makes use of the Extreme Gradient Boosting (XGBoost) method, random forest, and logistic regression. Data preprocessing includes removing irrelevant features, deleting duplicate records, processing missing values, and Z-score standardization. The results show that the XGBoost model performs best on the test set, with an recall rate, accuracy rate, precision rate and F1 score of 0.87, which is greater than logistic regression (0.774) and random forest (0.862). The AUC value of XGBoost is 0.95, and the classification ability is the best. Despite overfitting, XGBoost performs well in dealing with complex nonlinear relationships. Future research can optimize hyperparameters, expand data sets, and explore deep learning techniques to improve prediction accuracy and generalization ability further. This research offers evidence in favor of machine learning-based Parkinson's disease prediction, which is crucial for raising the bar for early detection.

Keywords: Parkinson, machine learning, logistic regression, random forest, XGBoost, data preprocessing, model evaluation.

1. Introduction

Parkinson's disease primarily affects the central nervous system and is a chronic neurological disease, causing movement problems such as muscle stiffness, bradykinesia and tremors. In addition, it may also lead to cognitive decline, emotional problems and other non-motor symptoms, seriously affecting the quality of life of patients. In the context of the current global aging trend, the incidence of neurodegenerative diseases, especially Parkinson 's disease, continues to rise, becoming an important factor that seriously threatens human health. According to statistics, the crude prevalence rate increased by about 74 % from 1990 to 2016 [1]. Parkinson's disease significantly impacts the patients' standard of living and imposes a substantial financial strain on families and society. Consequently, early and precise diagnosis of Parkinson's disease is crucial.

At present, Parkinson 's disease cannot be accurately predicted, and diagnosis usually depends on clinical evaluation and some diagnostic instruments. Neurological examination is one of the most basic methods. Doctors diagnose Parkinson 's disease through physical examination, imaging examination, functional evaluation, family history, and past diseases [2]. Although traditional methods play an important role in the diagnosis of Parkinson 's disease, these methods are highly dependent on the doctor 's judgment, the diagnosis results are susceptible to personal factors, and it is difficult to find early Parkinson 's disease [3]. With the rapid development of machine learning and data science, more and more research is devoted to using these technologies to predict and diagnose Parkinson 's disease. Compared with traditional methods, using machine learning to diagnose Parkinson 's disease is not affected by human factors, and has great advantages in dealing with multi-dimensional complex data. It can comprehensively analyze various factors and make early predictions of Parkinson 's disease. C Song et al. predicted Parkinson 's disease based on four feature selection algorithms and support vector machine (SVM). It turned out that alterations in brain areas including the lingual gyrus (LING) and superior frontal gyrus (SFGdor) may serve as biomarkers for Parkinson's disease. 92.24% categorization performance is the highest (specificity) [4]. IQ Khan predicts the severity and causes of Parkinson 's disease symptoms based on data mining methods. It

is found that the CART (J48 decision tree) algorithm performs best in the relationship that exists between the anticipated and actual values, with a correlation coefficient of 0.794 [5]. Although there are many algorithms, each method has its limitations, such as high demand for data quality and quantity or sensitivity to overfitting.

This paper attempts to investigate and compare the application of three machine learning algorithms: XGBoost, random forest, and logistic regression in predicting Parkinson's disease. By comparing the performance of these three algorithms on the actual Parkinson's disease data set, the most suitable model for Parkinson's disease prediction is identified. To achieve this goal, this study first performs data preprocessing, including data cleaning and feature selection, then constructs and trains each model, and evaluates their effects through multiple performance indicators including accuracy, recall, precision and F1 score. Finally, improvement strategies and future research directions are proposed based on the results obtained.

2. Data and methods

2.1. Data source and processing

The data set of this paper comes from the kaggle platform, which contains comprehensive health information of 2105 patients [6]. The data collection contains information on demographics, lifestyle choices, health history, clinical evaluations, cognitive and functional tests, symptoms, and diagnostic markers. The data set is of great value to the research of this paper, because it can help to explore the factors related to Parkinson's disease, develop predictive models, and conduct statistical analysis.

In order to preprocess the data, eliminate the interference factors, improve the data quality and the accuracy of the results, we must first remove irrelevant features, delete duplicate records, and fill or delete features with missing values. In addition, feature selection was performed through correlation analysis and feature importance score, and features with low correlation with the prediction of Parkinson's disease were eliminated. Finally, in order to eliminate the influence of different dimensions on the model, the data is standardized by Z-score to ensure the comparability between features. This series of elaborate data preparation steps lays a solid foundation for subsequent machine learning model training to ensure the validity and accuracy of the model.

2.2. Methods

This study used three machine learning algorithms to predict Parkinson's disease, including logistic regression, random forest and XGBoost. Each algorithm's concept, principle and characteristics are introduced in detail below.

2.2.1 Logistic regression.

It is a popular linear model for classification issues. The basic idea behind it is to use the linear relationship between the target variable's characteristics and the data set's characteristics to predict the target variable's probability [7]. Firstly, logistic regression linearly combines the input features to obtain the output of the linear model:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \quad (1)$$

Among them, β_0 is the bias term, β_i is the coefficient of each feature and x_i is the input feature. What the linear model produces is mapped to the probability value through the logic function (Sigmoid function):

$$P(y = 1|x) = \frac{1}{1+e^{-z}} \quad (2)$$

Where $P(y = 1|x)$ represents the probability of an event (category 1) when the input feature is x . The Maximum Likelihood Estimation (MLE) method is used to solve the parameters β , make the probability value predicted by the model closest to the actual value. In the specific implementation, the gradient descent algorithm is commonly used to optimize the parameters.

The probability of an occurrence may be simple to comprehend from the logistic regression model's output, which makes the model easy to understand and explain. The coefficients of each feature represent the degree of influence on the results. This intuitive interpretation helps to understand the model's decision-making process and has important reference value in practical applications. In terms of computational efficiency, logistic regression performs well and is especially suitable for processing large-scale data sets. Due to its linear model-based structure, logistic regression has relatively low computational complexity and can quickly process many samples and features. This feature makes it particularly suitable for application scenarios that require real-time or near-real-time prediction. Logistic regression can also effectively deal with the problem of multicollinearity between features through regularization methods (such as L1 and L2 regularization). The output of logistic regression is a probability value, which can flexibly adjust the classification threshold according to the needs of specific application scenarios. This is very useful in many practical problems, such as credit scoring, disease prediction, and other application scenarios, which require further decisions based on probability values. Despite the fact that logistic regression presumes a linear connection between the target variables and features, introducing feature engineering, such as polynomial features or interactive features, can expand the application scope of logistic regression and improve its performance on nonlinear problems.

2.2.2 Random forest.

This primary applications of this ensemble learning technique are in the areas of regression and classification [8]. The basic idea is to improve the model's prediction performance and generalization ability by generating several decision trees and merging the results of each one. The specific principle is as follows:

Firstly, random forest extracts multiple sub-sample sets from the original training set by Bootstrap method. For each sub-sample set, a decision tree is trained. Each decision tree will randomly select a part of features at each node for splitting during training, thus broadening the model's range of possibilities. Next, random forests vote on each decision tree's prediction findings to determine the final classification result in the classification challenge. In the regression task, random forest determines the final prediction result by averaging each decision tree's predicted output. Through the above process, random forest combines the advantages of multiple decision trees, effectively lowers the possibility of a single model becoming overfit, and improves the stability and prediction performance of the model.

Random forest has high accuracy and robustness. Because random forest reduces the over-fitting problem that may occur in a single decision tree by combining the prediction results of multiple decision trees, it often performs well in practical applications [9]. Secondly, Numerous attributes and high-dimensional data can be handled by random forests. Since each decision tree uses only a part of features during training, random forests will not have overfitting problems when dealing with high-dimensional data. This makes it particularly suitable for data sets containing a large number of features. Random forests can also handle missing values and maintain high prediction performance. By randomly selecting sub-sample sets and features, random forests are able to handle missing values in the data in an efficient manner without adversely impacting the overall performance of the model. The random forest also includes an integrated feature importance assessment tool [10]. During the training process, random forest calculates the importance of each feature in different decision trees, thus providing an effective feature selection method. This is very useful for understanding important data features and improving the model's interpretability. The random forest model is relatively simple, easy to implement, and has high parallelization potential. Since each decision tree can be trained and predicted independently, random forest is very suitable for running in multi-core processors or distributed computing environments, greatly improving computational efficiency.

2.2.3 XGBoost.

XGBoost is an ensemble learning technique mainly used for classification and regression tasks. It is based on the Gradient Boosted Decision Tree (GBDT) [11]. The principal concept is to reduce the

error by gradually adding a new decision tree based on the existing model. The specific principle is as follows : First, XGBoost initializes a simple model (usually predicting that all samples are average). It then guides the generation of new decision trees by calculating residuals (the discrepancy between the present model's actual and anticipated values). With each iteration, the XGBoost algorithm adds a new decision tree that aims to accurately capture the residuals left by the previous round. Then, XGBoost updates the model's predicted value by linearly combining the output of all decision trees. Different from the traditional gradient boosting method, regularization is a phrase that XGBoost adds to the training process to regulate the complexity of the model and avoid overfitting. XGBoost also uses distributed computing and parallel processing technology to greatly improve training speed and computational efficiency.

XGBoost has a very high prediction performance. By gradually fitting the residuals, XGBoost can capture complex patterns in the data, thus performing well in many practical applications. Secondly, XGBoost has strong anti-overfitting ability. By introducing a regularization term into the training process, XGBoost can effectively reduce the model's complexity to avoid the training set becoming overfit [12]. This makes XGBoost perform better for handling complex high-dimensional data. XGBoost also has efficient computing performance [13]. Using distributed computing and parallel processing technology, XGBoost can process large-scale data sets in a short time, which is particularly important in application scenarios that require fast response. GBoost provides a wealth of parameter adjustment options, allowing users to flexibly adjust the parameters of the model according to the needs of specific problems, thereby further improving the model's prediction performance.

Compared with logistic regression, XGBoost can handle more complex nonlinear relationships. A linear relationship between the features and the target variable is the foundation of logistic regression. which is better when dealing with simple problems, but may not perform well in complex problems. By iteratively training multiple decision trees, XGBoost can capture the nonlinear relationship in the data, thereby improving the model's prediction ability.

Compared with random forest, XGBoost is more efficient in processing data and usually has higher prediction performance. Random forest builds several independent decision trees to increase the model's stability and accuracy, and XGBoost combines the output of each decision tree by gradually fitting the residuals so as to better reduce the error. In addition, the regularization mechanism of XGBoost also makes it more effective than random forests in preventing overfitting.

The characteristics of XGBoost include its high efficiency because it has extremely high computational efficiency and predictive performance when dealing with large-scale data. It has strong flexibility, supports multiple objective and custom loss functions, and has strong adaptability. XGBoost also controls the complexity of the model through regularization terms, thereby effectively preventing overfitting.

In summary, logistic regression, random forest and XGBoost have their own unique advantages and applicable scenarios. Because of its simplicity and high efficiency, logistic regression is suitable for problems with obvious linear relationships. Random forest is suitable for high-dimensional and complex data sets because of its high precision and robustness. With its high efficiency and flexibility, XGBoost is especially suitable for modeling large-scale data and complex nonlinear relationships. In this study, these three algorithms will be applied to model training and comparison to obtain the best Parkinson 's disease prediction model.

2.3. Evaluation index

This article uses four metrics to evaluate the performance of XGBoost, random forest, and logistic regression models in identifying Parkinson's disease. Accuracy, recall, precision, and F1 scores are some of these metrics. The classification ability of the model can be fully assessed using these indicators. Accuracy can be defined as the proportion of correctly predicted samples to all samples. The accuracy rate is calculated using the following formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (3)$$

This instance uses the notation TP for true class, TN for true negative class, FP for false positive class, and FN for false negative class to denote the various classes. The accuracy of the model is measured in a range of values from 0 to 1, where a higher value indicates more overall accuracy. Although this indicator can show the model's overall success in classifying samples across all samples, it may not accurately represent the model's performance when the sample category is unbalanced, meaning that the proportion of positive and negative samples differs significantly.

Recall is the proportion of samples among all true positive samples that the model correctly predicted as positive. The formula for calculating it is:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

The ability of Recall to identify positive samples can be represented by this metric. Especially in scenarios such as disease diagnosis that need to minimize missed diagnosis, the recall rate is an important evaluation indicator.

The proportion of samples that are truly positive among all samples predicted by the model to be positive is called precision. This is how it is calculated:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5)$$

The value of the precision ranges from 0 to 1, and the higher the value, the higher the proportion of the actual positive class in the samples predicted by the model as positive class. This indicator can reflect the prediction accuracy of the model. Precision is a key indicator when misdiagnosis or false prediction needs to be reduced.

The F1 score is a complete evaluation of the precision and recall of the model, calculated as the harmonic mean of the two variables. Its calculation formula is:

$$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

The value range of F1 score is between 0 and 1, and the higher the value is, the better the model performs in balancing accuracy and recall. This index can comprehensively evaluate the precision and recall rate of the model, especially for scenarios with unbalanced sample categories, and can more comprehensively reflect the classification performance of the model.

Through these four evaluation indicators, the performance of logistic regression, random forest and XGBoost models in Parkinson 's disease diagnosis tasks can be comprehensively measured, providing a reliable basis for comparison and analysis.

3. Results

To evaluate the model, the model dataset is divided into a training set and a test set, where 70% of the dataset is designated as the training set. The percentage of accurately predicted samples among all samples is called accuracy. Recall indicates the proportion of true positive cases that are correctly identified, while accuracy indicates the proportion of expected positive cases that are actually positive. Higher values for these three metrics signify better model performance. However, recall and precision often have an inverse relationship in practice. To effectively combine recall and precision, the harmonic mean known as the F1 score is used, resulting in the findings presented in Table 1.

Table 1. Three Model Assessment Tables

	Accuracy	Recall	Precision	F1
Logical regression training set	0.845	0.845	0.844	0.844
Logical regression test set	0.774	0.774	0.772	0.772
Random forest training set	0.978	0.978	0.978	0.977
Random forest test set	0.862	0.862	0.862	0.862
XGBoost training set	1	1	1	1
XGBoost test set	0.87	0.87	0.87	0.87

As shown in Table 1, the logistic regression model performs the worst on the test set, with an accuracy of 0.774, a recall rate of 0.774, an accuracy of 0.772, and an F1 score of 0.772. In contrast, random forest and XGBoost perform better on the test set, especially XGBoost 's indicators have reached 0.87. The four metrics of the random forest model on the test set are 0.862, which is better than the logistic regression model. The performance of the XGBoost model on the test set is slightly better than that of the random forest, with an accuracy rate, recall rate, precision rate and F1 score of 0.87. In general, XGBoost has the highest accuracy, recall, precision and F1 value on the test set, indicating that it has the strongest predictive ability and the best effect on new data.

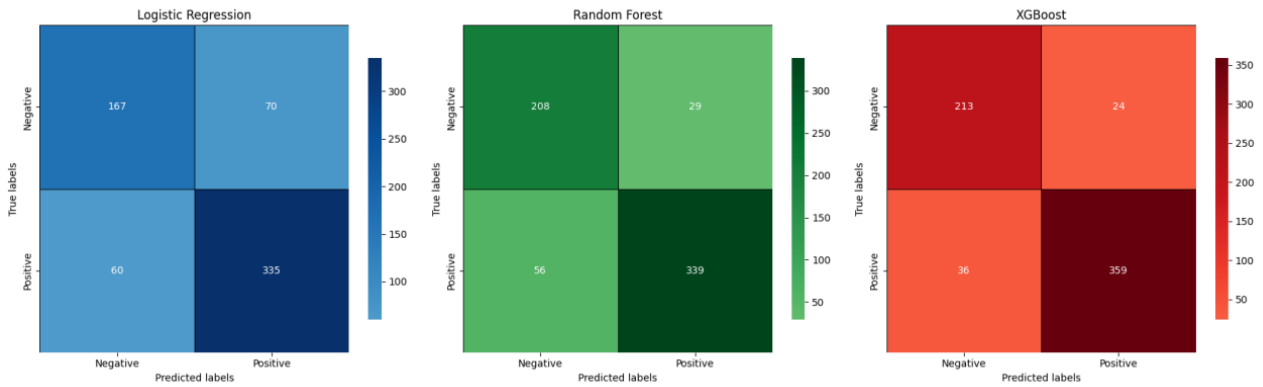
**Figure 1.** Comparison of confusion matrices

Figure 1 demonstrates the three machine learning models' performance in classification tasks as indicated by the confusion matrix. The confusion matrix provides an intuitive view for evaluating the performance of the classification model by showing the correspondence between the actual category and the predicted category. 0.0 and 1.0 indicate the prediction and actual classification of negative and positive classes, respectively.

The confusion matrix of the logistic regression model shows a certain number of true negative classes and true classes, which are 172 and 317 respectively. However, the model also produces 79 false-positive classes and 64 false negative classes, indicating that the model has a certain degree of misjudgment in distinguishing these two categories. The random forest model showed better classification ability, the true negative class increased to 200, the true class was 345, and the number of false positive and false negative classes decreased to 51 and 36. This reflects the advantages of random forests in reducing misjudgments, showing higher classification accuracy and better category sensitivity. Out of the three models, the XGBoost model has the best performance. It has a true negative class of 203, a true class of up to 347, and further reductions to false positive and false negative classes of 48 and 34. This result highlights the powerful ability of XGBoost in improving classification accuracy and reduce misclassification, especially in distinguishing positive and negative classes.

The XGBoost model shows the best overall performance in this group of comparisons, followed by random forest. Although logistic regression performs well, it has a high misjudgment rate. These differences indicate that there may be significant differences in the performance of different algorithms when dealing with the same data set, and selecting the appropriate model is crucial for achieving high-accuracy classification. Through the visualization of this confusion matrix, researchers can more clearly identify the advantages and disadvantages of each model in specific tasks, and provide a basis for further model selection and optimization.

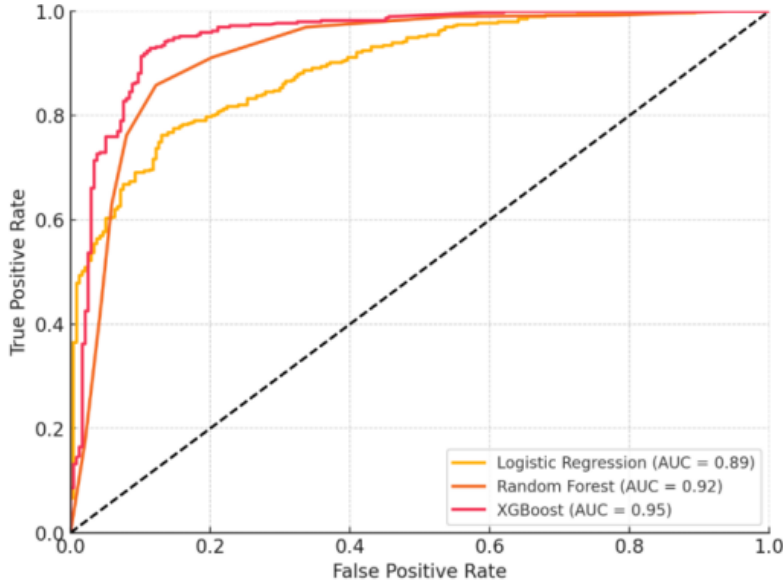


Figure 2. ROC curve of three models

As shown in Figure 2, Under different threshold settings, a comparison of the model's true positive rate and false positive rate was evaluated by the ROC curve. The Area Under the Curve (AUC) value measures the overall performance of the model in classification. The better the model performance, the higher the AUC value.

The AUC value of logistic regression was 0.89. The curve is nearer to the top left corner, showing a higher classification effect but is the lowest between the three models. The logistic regression model performs well in the lower false positive rate region, but as the false positive rate increases, the increase in its true rate gradually slows down. The AUC value of random forest was 0.92. This curve is closer to the top left corner than logistic regression, indicating that random forests can achieve higher true rates at the same false positive rate level, and thus perform better. XGBoost has the highest AUC value of 0.95. This curve is almost close to the left and upper boundaries of the chart, showing excellent classification ability, and maintains a low false positive rate and a high true rate under almost every threshold configuration.

4. Discussion

This study has achieved certain results in predicting Parkinson's disease using logistic regression, random forest and XGBoost models. However, in terms of data sets, the data set of this experiment is small, with only 2105 samples, which may limit the capacity of the model to be generalized, and the source of the data set is single, which may not be widely representative, influencing the model's suitability for use with various populations. In the future, the hyperparameters of the model can be further adjusted and optimized, and higher-level parameter optimization methods (such as evolutionary algorithms, Bayesian optimization, etc.) can be applied to enhance the performance of the model. Furthermore, we can also try to integrate the outcomes of several models' predictions to combine the advantages of each model to further enhance the prediction effect.

Moreover, the model employed in this study showed good performance on the training set but comparatively poor performance on the test set. The main reason may be overfitting. On the training

set, the XGBoost model performs at 1.0, which further demonstrates this situation. The possible causes of overfitting include high model complexity, insufficient data set size and lack of regularization. The XGBoost model has high flexibility and complexity, and may over-fit the noise in the training data during the method of training. In addition, the amount of training data is small, which makes it easy for the model to perform well on training data but not well on test data. Regularization is an important means to prevent overfitting. The generalization ability of the model can be enhanced by adding a penalty term to limit the complexity of the model.

5. Conclusion

This study examined the efficacy of three machine learning algorithms in predicting Parkinson's disease: XGBoost, random forest, and logistic regression. To guarantee the model's efficacy and precision, a dataset of 2,105 patients on the Kaggle platform was preprocessed and z-score standardized.

It was found that the three models performed differently on the training set and the test set. The F1 score, recall, accuracy, and precision of logistic regression on the test set are all 0.774, which is the worst among the three models. The indexes of random forest were 0.862, while the XGBoost model performed best on the test set, and the indexes reached 0.87. In addition, from the confusion matrix and ROC curve, It is clear that the XGBoost model functions the best. in reducing misclassification and improving the overall classification ability, and its AUC value reaches 0.95, which is significantly higher than logistic regression and random forest.

Although the XGBoost model performs well in predicting Parkinson's disease, there is also a certain over-fitting phenomenon. Future research can enhance the model's performance by optimizing hyperparameters, introducing regularization, expanding the scale of data sets, adopting more advanced parameter optimization techniques, and trying to integrate the prediction results of multiple models. This study highlights machine learning's potential in Parkinson's disease prediction, especially the advantages of XGBoost model. Future research directions also include exploring the application of deep learning technology, and combining real-time monitoring data to achieve dynamic prediction and early warning of Parkinson 's disease.

Reference

- [1] Y. Ben-Shlomo, S. Darweesh, J. Llibre-Guerra, C. Marras, M. San Luciano, & C. Tanner. The epidemiology of Parkinson's disease. *The Lancet*, 403(10423), 2024, 283-292.
- [2] A. Berardelli, G. K. Wenning, A. Antonini, D. Berg, B. R. Bloem, V. Bonifati, & W. Poewe. EFNS/MDS-ES recommendations for the diagnosis of Parkinson's disease. *European Journal of Neurology*, 20(1), 2013,16-34.
- [3] C. W. Olanow, M. B. Stern, & K. Sethi. The scientific and clinical basis for the treatment of Parkinson disease 2009. *Neurology*, 72(21 Supplement 4), S1-S136.
- [4] C. Song, W. Zhao, H. Jiang, & Y. Tang. Stability evaluation of brain changes in Parkinson's disease based on machine learning. *Frontiers in Computational Neuroscience* 2021.
- [5] I. Q. Khan. Simultaneous prediction of symptom severity and cause in data from a test battery for Parkinson patients, using machine learning methods. *Technology* 2009.
- [6] R. Elkharaoua. (2024). Parkinson's Disease Dataset Analysis. Kaggle. Retrieved from <https://www.kaggle.com/datasets/rabieelkharoua/parkinsons-disease-dataset-analysis>
- [7] D. W. Hosmer, S. Lemeshow, & R. X. Sturdivant. *Applied Logistic Regression* (3rd ed.). John Wiley & Sons 2013.
- [8] L. Breiman. Random forests. *Machine Learning*, 45(1), 2001, 5-32.
- [9] A. Liaw, & M. Wiener. Classification and regression by randomForest. *R News*, 2(3), 2002,18-22.
- [10] D. R. Cutler, T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, & J. J. Lawler. Random forests for classification in ecology. *Ecology*, 88(11),2007, 2783-2792.

- [11] T. Chen, & C. Guestrin. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ,2016, 785-794.
- [12] J. Zhang, Y. Zhao, & X. Chen. Improved XGBoost algorithm based on weighted feature selection and parameter optimization. IEEE Access, 8, 2020, 132128-132139.
- [13] S. M. Lundberg, & S.-I. Lee. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 2017,30