Predicting Parkinson's Disease Progression with Random Forests

Shuxin Qian
Math College, South China University of Technology, Guangzhou, 510640, China
nyinds@outlook.com

ABSTRACT

This paper proposes a prediction method for Parkinson's disease (PD) progression based on the random forest model. Because a large number of proteins are involved in the development of PD, protein data can be used to help diagnose PD progression. In this work, approximately 2,600 disease data (MDS-UPDRS) and protein data from 248 patients are used to establish a random forest model for predicting PD progression. The initial step involved rigorous data pre-processing to address missing values, and the dataset dimension is reduced to 500 by the Principal Component Analysis (PCA) method. The random forest model used in the experiment uses bagging to generate decision trees, which contain a total of 100 decision trees. Bagging can generate a large number of subsets for model building, which can greatly improve model generalization performance, while decision trees are nonlinear predictive models with superior performance. After building the predictive model, we analyse the residuals and explain the model using metrics such as SHAP and feature contribution. In addition, we introduce a Multiple Linear Regression (MLR) prediction model and conduct a comparative analysis between it and the random forest prediction model. To evaluate the model's performance, we employed two critical indicators, the Mean Squared Error (MSE) and the Symmetric Mean Absolute Percentage Error (SMAPE). Our findings indicated an MSE of 27.54 and an SMAPE of 69.50 for the random forest model, while The MSE and SMAPE of the MLR model are 55.81 and 93.73. After comparison, the model has better performance than the multiple linear regression model (MLR) and has a certain application value. The model's potential impact includes enhancing PD diagnosis accuracy, enabling early intervention and treatment, thereby improving patients' quality of life and reducing healthcare costs. It can optimize medical resource allocation, leading to greater efficiency in the healthcare system.

CCS CONCEPTS

• **Applied computing** → Life and medical sciences; Health care information systems.

KEYWORDS

Parkinson's Disease, Random forest, Protein

BIC 2024, January 26–28, 2024, Beijing, China

ACM ISBN 979-8-4007-1664-5/24/01
https://doi.org/10.1145/3665689.3665722

ACM Reference Format:
Shuxin Qian. 2024. Predicting Parkinson's Disease Progression with Random Forests. In *2024 4th International Conference on Bioinformatics and Intelligent Computing (BIC 2024), January 26–28, 2024, Beijing, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3665689.3665722

1 INTRODUCTION

PD is a neurodegenerative disorder associated with age. Its symptoms include stiffness, tremors, unsteady gait, and even movement disorders [1]. In 2016, approximately 6 million people worldwide were affected by PD, and more are expected to suffer from PD in the future [2]. And as PD progresses, its impact on individuals becomes increasingly profound, leading to a declining quality of life for patients [3]. But to treat the disease, patients have to incur significant medical expenses. According to current research, the development of PD is closely linked to abnormalities in certain proteins and peptides within the human body.

Current research suggests that PD is likely caused by neuronal necrosis in the substantia nigra. In addition, misfolding of some proteins plays a key role in the pathogenesis of PD. Proteins such as DJ-1 protein, MIA, CPR and albumin are all closely related to PD [4, 5]. Therefore, it can be known that the concentration of proteins associated with PD in Parkinson's patients may be different from that of normal people. Furthermore, it is likely that the concentrations of these proteins will continue to change as PD progresses.

PD exhibits a latent phase with no obvious symptoms initially [6]. Besides, its symptoms are similar to other neurological disorders. So, clinical diagnosis is complicated. Since there are proteins and peptides in Parkinson's patients that are clearly related to the disease, we can collect these protein content data and the patient's condition data, and then use machine learning methods to predict the progression of PD.

Nowadays, doctors are increasingly using machine learning algorithms to diagnose patients, and these algorithms can make good use of the patient data collected by doctors to make judgments. These judgments are often accurate and effective, sometimes even better than those of some experienced doctors. Therefore, the study uses the random forest algorithm to predict the progression of PD.

The random forest algorithm has been applied to a considerable extent in the medical field and has achieved considerable results. Such as, automatic detection of Alzheimer's disease and breast cancer diagnosis [7, 8].

2 DEVELOPMENT OF EXISTING TECHNOLOGIES AND WORK

In fact, Since the initial discovery of Parkinson's disease, researchers and medical professionals have been tirelessly working to identify methods for early detection, prediction, and treatment. Recent developments in the field of machine learning and deep learning,

combined with patient data, have led to the creation of predictive systems for early Parkinson's disease. These systems have the potential to revolutionize early intervention strategies and enhance the overall quality of life for individuals living with this condition. Moreover, machine learning techniques have been harnessed to construct models that can predict changes in the quality of life for Parkinson's disease patients. Such models can provide invaluable insights into the progression of the disease and offer healthcare providers the opportunity to tailor interventions for each patient's specific needs. Furthermore, recent research endeavors have harnessed the power of deep learning techniques in the analysis of retinal images, offering a novel approach to assessing the severity of neurological dysfunction in Parkinson's disease [9].

In addition to these technical advancements, there is a cohort of scholars dedicated to addressing the social aspects of Parkinson's disease. They have taken on the noble task of disseminating knowledge about Parkinson's disease to underserved, remote rural areas. Their efforts are not only informative but also represent a holistic approach to the disease's management and awareness.

What's more, we must acknowledge the valuable contributions from scholars in related fields who have advanced our understanding of Parkinson's disease. The pioneering efforts and accomplishments of these scholars have played a pivotal role in shaping the landscape of Parkinson's disease research and have profoundly inspired the inception of our own experiment.

3 DATA PRE-PROCESSING

In our experiment, we utilized a random forest algorithm to forecast the progression of PD. The dataset for this study was protein abundance data sourced from Kaggle (https://www.kaggle.com/competitions/amp-parkinsons-disease-progression-prediction/data). The data set consists of three parts, namely clinical data, proteins and peptides. clinical data includes Parkinson's patient ID, time of visit relative to first test, and test score. The test score is based on the MDS-UPDRS to test and get a specific score. MDS-UPDRS is used to evaluate the severity of PD symptoms. It mainly tests four parts, namely non-sports experience in daily life, sports experience in sports life, motion detection and sports complications [10]. The higher the score, the more severe the symptoms.

The proteins dataset contains the Parkinson's patient ID, the time of visit (relative to the first visit) and the detected proteins and their concentrations. The peptides dataset contains the Parkinson's patient ID, the time of visit (relative to the first visit) and the detected peptides and their concentrations.

The experiment needs to predict the patient's MDS-UPDRS score based on the Parkinson's patient's ID, visit time, and detected protein and peptide content, which are updrs_1, updrs_2, updrs_3 and updrs_4 respectively. So, we use the three datasets to make a new dataset called Parkinson-dataset.

First, we import Parkinson's patient ID and visit time into Parkinson-dataset. Then based on two attributes, we extract the corresponding attribute value in UniProt as the attribute of Parkinson-dataset, and then extract the NPX attribute value in the same row corresponding to the attribute value in UniProt and assign it to the corresponding UniProt attribute in Parkinson-dataset. Then, we do a similar operation on the peptide dataset. Convert the attribute values of Peptide and Peptide Abundance in the peptide dataset into new attributes and corresponding new attribute values in the Parkinson-dataset, respectively.

Then we check the Parkinson-dataset and find that there are many NULLs in it, which indicate that certain experimental data are missing. Based on the situation, for simple calculations, we use the average value of the corresponding attribute to fill the NULL.

In addition, we find that the data size of different proteins and peptide attributes in the Parkinson-dataset varies greatly. For example, the value of protein O00391 is from 620.87 to 19799.8, while the value of protein O00533 is from 77941.7 to 1354570.0. This should be because the functions of different proteins in the human body are different, so the amount of them in the human body is different. The gap of amount is likely to affect the accuracy of the random forest algorithm. For this reason, a normalization method is used to transform the attributes so that the impact of differences between different proteins can be eliminated [11].

In addition, after establishing the random forest model, it is necessary to test the generalization performance of the model. We divide the data set into a test set and a training set. The training set is employed to refine the model, while the test set serves to assess the model's performance.

4 DATA EXPLORATION ANALYSIS

First, we look at the dataset, including the median, maximum, minimum, mean, and standard deviation (Table 1). We can find that the median, mean and standard deviation of O00391 is 11546.40, 11641.26 and 2817.00. Besides, we can know statistics for other attributes. We can find that there is a gap between the statistics of these different attributes. Differences in these protein statistics may indicate that different proteins have varying associations with PD. So, we need to explore this further.

After normalizing the Parkinson's dataset, we took out the eight protein or peptide content attributes and drew the corresponding box plots (Figure 1). These attributes are LTASAPGYLAITK, P01034, EQWPQC(UniMod_4) PTIK, ITGYIIK, SLHTLFGDK, P05090, GLSAEPGWQAK and P02792. We find that most of the outliers are distributed above the box plot. The distribution of these outliers is likely to have a certain relationship with PD, which may help us better understand PD.

We create a line graph of PD progression over time based on months and patient MDS-UPDRS scores in the Parkinson-dataset. Although updrs_1, updrs_2 and updrs_3 is constantly fluctuating, they generally increase over time. And there is a high correlation between updrs_1 and score updrs_2. Although the updrs_4 curve changes slowly, we can also see that it generally increases over time. Obviously, PD develops over time, although there will be slight relapses in the middle.

We then sample two patients' data from the Parkinson's dataset and plot two-line graphs (Figure 2 and Figure 3) of their UPDRS scores over time. We can see that the patient's score fluctuates frequently, and his ID is 7886. It shows that the patient has received good treatment and his condition is under certain control. However, judging from the overall score, his condition is getting worse.

The other line chart is about the patient, whose ID is 942. From the line chart, we can see that the patient's updrs_1, updrs_2 and

Table 1: Summary of protein data

	O00391	O00533	...	YYTYLIMNK	YYWGGQYTWDMAK
count	764.000000	1.112000e+03	...	1030.000000	865.00000
mean	11641.264435	5.111649e+05	...	47068.709311	21072.04823
std	2817.003530	2.357357e+05	...	13689.667117	10360.59380
min	873.778000	5.971820e+04	...	6362.490000	868.90300
25%	9736.857500	3.490590e+05	...	37752.375000	14249.90000
50%	11546.400000	4.834425e+05	...	45503.150000	20390.90000
75%	13383.025000	6.485572e+05	...	54748.350000	27031.90000
max	21361.800000	1.806980e+06	...	107220.000000	70020.80000

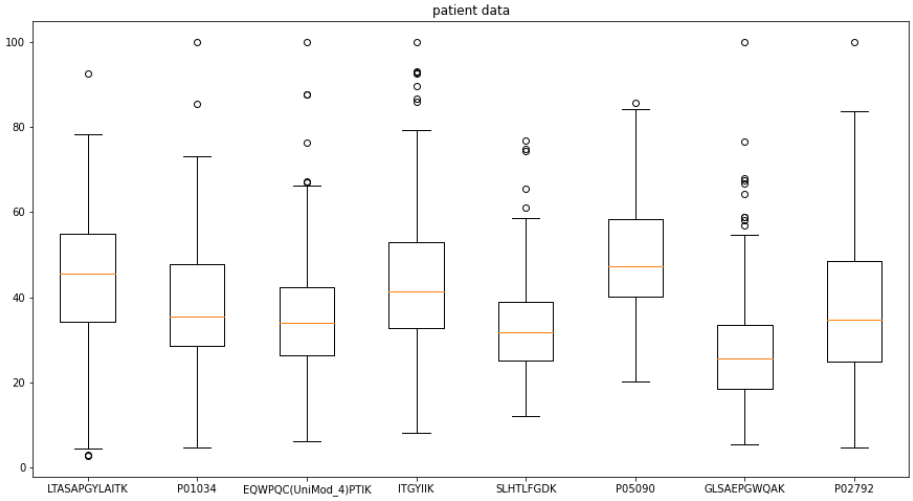

Figure 1: the box of proteins

updrs_4 is all relatively low, all below 10 points, and the fluctuations are relatively small. However, we can see that the patient's updrs_3 fluctuates greatly, and the upward trend is obvious.

Judging from the UPDRS scores changes of patients in two departments, PD is difficult to control long-term and difficult to predict using traditional methods. For this reason, it is very meaningful to use machine learning methods to predict PD.

In addition, we also extract a part of the protein content data of the patient whose ID is 7886 in certain months and make a histogram (Figure 4). As can be seen from these figures, the levels of certain proteins change as PD progresses. For example, protein contents of P00450, P00734 and P00738 increase over time generally. In addition, some protein content fluctuates continuously over time. Such as P01833. It shows that these protein levels are likely to have a higher correlation with the progression of PD.

Of course, there are some proteins that remain essentially unchanged. Such as P01034. These proteins may not be associated with Parkinson's. In future studies, we need to remove the interference of the possibly irrelevant proteins.

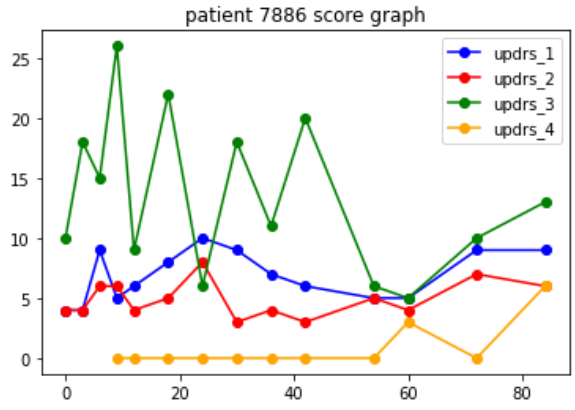

Figure 2: Patient 7886 disease progression

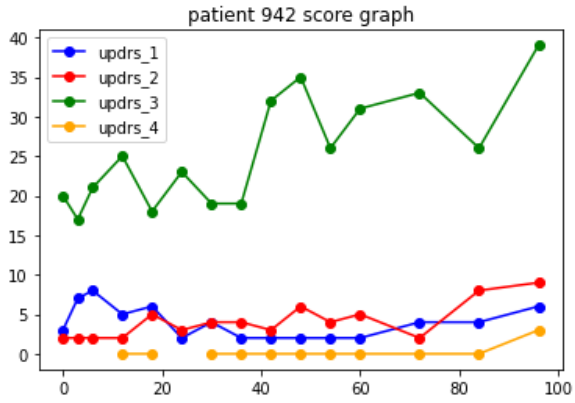

Figure 3: Patient 942 disease progression

5 METHODS ANALYSIS

5.1 Method Overview

5.1.1 Random Forest Model. Random Forest, currently the strongest and most commonly used supervised algorithm, determines the output of the overall model by combining the outputs of multiple weaker decision tree models [12]. The more decision trees its model contains, the higher its effectiveness and stability. Due to the characteristics of the algorithm itself, random forest has good generalization performance, that is, it can achieve better fitting results on real data.

First of all, the random forest algorithm is composed of a series of decision trees [13]. The output of random forest regression is determined by the average of the common output of all decision trees. A decision tree is composed of root node, leaf nodes, and decision nodes. The determination of these nodes is usually determined by the Gini index.

$$Gini\ Index = 1 - \sum_{i=1}^{n} P_i^2 \tag{1}$$

p_i represents the overall proportion of data points at the $i-th$dividing point.

5.1.2 . Multiple Linear Regression Model (MLR).. In order to better demonstrate the effect of the random forest model, we use MLR for comparison [14, 15]. The multiple regression model is shown below.

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \tag{2}$$

$$Q\left(\beta_0 \beta_1 \ldots \beta_p\right) = \sum_i^n \left(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip}\right)^2 \tag{3}$$

The loss function of the model is the sum of squared residuals. $y$is the predicted value.β_i is the regression coefficient, and ε is a random interference item.

5.2 Model Fitting

In the Parkinson's dataset, we can see the month of the patient visit, the protein content, the peptide content of the patient and UPDRS scores. Our experience shows that for these proteins and peptides, there are some missing in the peptide data.

Besides, some protein content data of many patients are missing, and the experiment uses the average value to fill in. The filling method has a better effect and less calculation. Perform PCA dimensionality reduction on the Parkinson-dataset to filter out noise and redundant information and generate a dataset with less correlation after dimensionality reduction [16]. This approach can significantly decrease the computational time and enhance the model's generalization performance.

The entire PCA dimensionality reduction process can be expressed by the following formula.

$$X^* = XP \tag{4}$$

$X$is the original Parkinson-dataset matrix and $X*$ is the data matrix dimensionally reduced. p is a matrix composed of d orthogonal eigenvectors of X, where d is equal to the dimension of $X*$.

The Parkinson's dataset used in the experiment is a high-dimensional dataset with a total of 1196 attributes. Even if the dimensionality is reduced through PCA, there are still 500 attributes. The random forest model can handle the high-dimensional data very well. In addition, our Parkinson's dataset has a lot of missing information, while the random forest model is not sensitive to missing data and can guarantee the model effect.

Experiments are conducted on the dataset using the n-fold crossover method. The dataset is then split into two parts: a training set and a test set, with an 8:2 ratio.

Since there are four UPDRS scores, we need to perform random forest modeling using each score separately with the new attributes in the dataset naked by PCA.

After importing the train set into a random forest model, the model uses bagging to generate subsets of the original dataset and distribute the subsets to the decision trees in the model respectively (Figure 5). These decision trees are trained using these subsets. When the decision trees are trained, it indicates that the random forest model of the experiment is trained too.

5.3 Model Analysis

Nowadays, random forest models use bagging as an operation [17]. Bagging is the use of random sampling with replacement to generate subsets (Figure 6). Typically, the sample size of the subset is equivalent to the size of the original dataset. Due to the random sampling with replacement to generate itself, there may be duplicate samples in the subset or some samples from the original dataset that do not appear. The process of bagging usually generates multiple subsets. The number of subsets utilized in bagging is generally equivalent to the number of decision trees utilized in a random forest. The entire process of generating subsets through bagging is illustrated in the Figure 6.

Since The data are generated by random sampling, they are independent of each other.

Since the decision trees in a random forest model are trained on different subsets, they can learn different aspects and noises of the original data. Therefore, the predictions made by the decision trees are not significantly correlated. Therefore, the random forest model based on the decision trees have lower variance, significantly improve stability, and greatly enhance generalization performance.

The random forest regression model is interpretable. We can judge how the model makes decisions by looking at the decision-making process of each decision tree in the model.

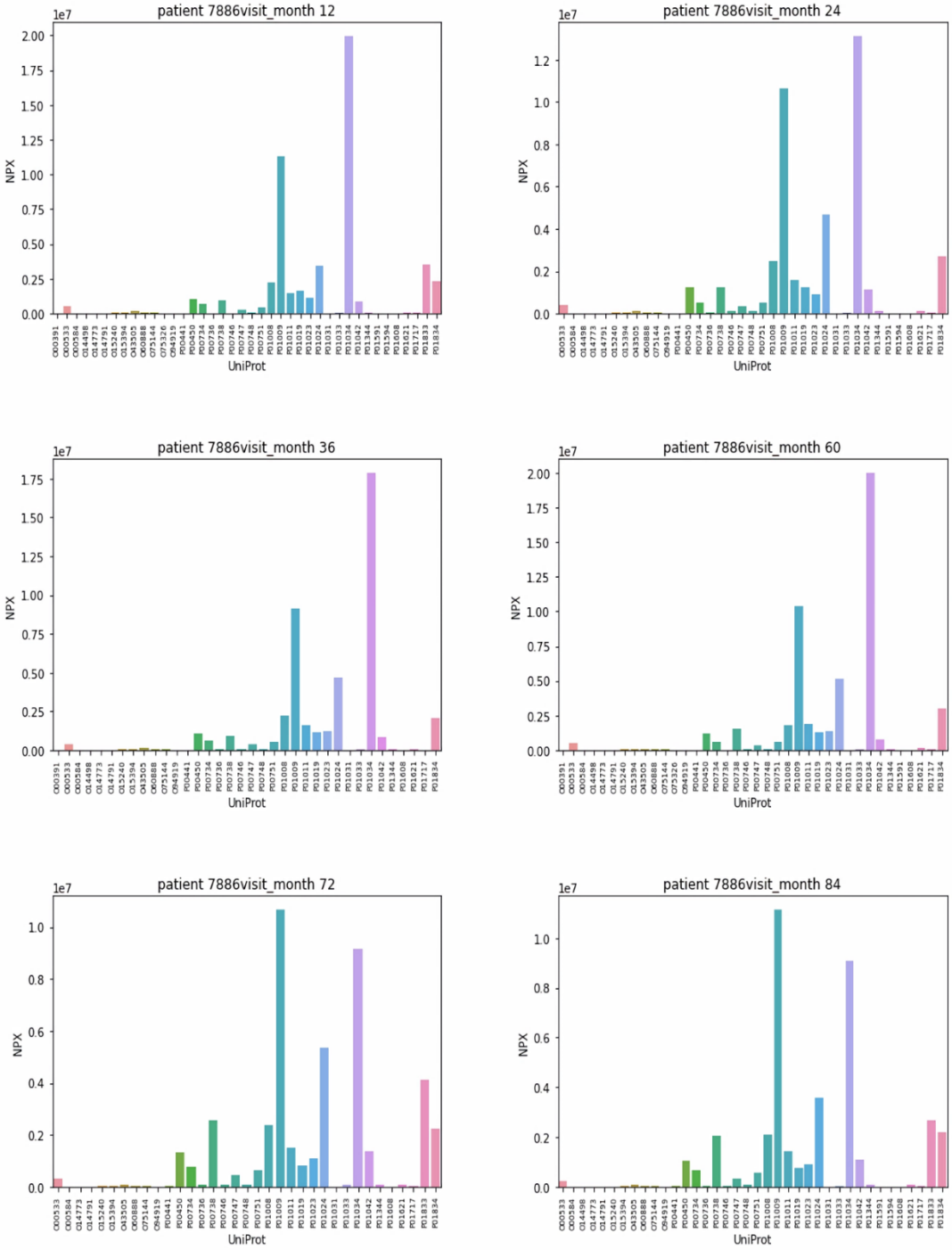

Figure 4: Protein changes in Parkinson's disease patient 7886

The performance of a random forest regression model is closely correlated with the number of decision trees incorporated in it. The Figure 7 is an image of the performance of a random forest model containing different numbers of decision trees trained by using PCA-processed data attributes and patient updrs_4. It can be seen that the performance of random forest on the training set is much better than that on the test set.

Furthermore, our findings indicate that there is a rapid improvement in the performance of the random forest model when the number of decision trees ranges from 0 to 10. However, beyond

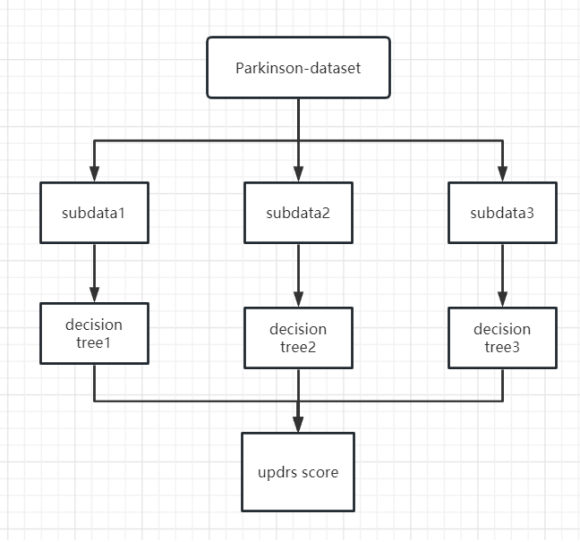

Figure 5: The random forest model

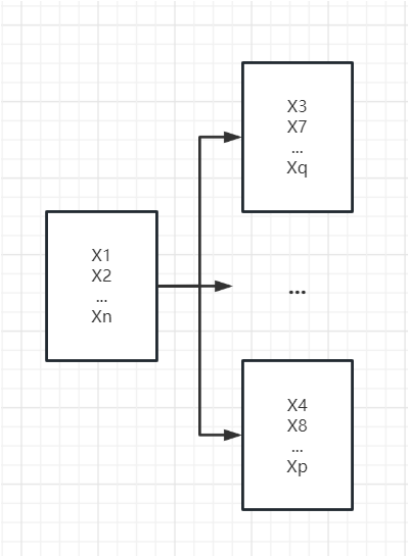

Figure 6: The process of bagging

this range, the improvement in performance becomes negligible. The Figure 7 shows that with the increase in the number of decision trees, the time for model training will continue to increase. Consider the trade-off between random forest model performance and training time. We set the number of decision trees to 50.

5.4 Model Diagnostics

The sum of random forests is very interpretable. The random forest model can detect the importance score of each feature [18]. The Figure 8 below is an image of the random forest model on the patient updrs_1 score and the importance scores of each attribute of the PCA-processed data set. We can see that the importance

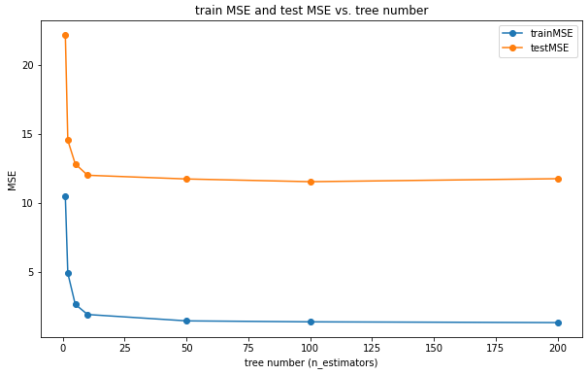

Figure 7: The relationship between MSE decrease and the number of trees

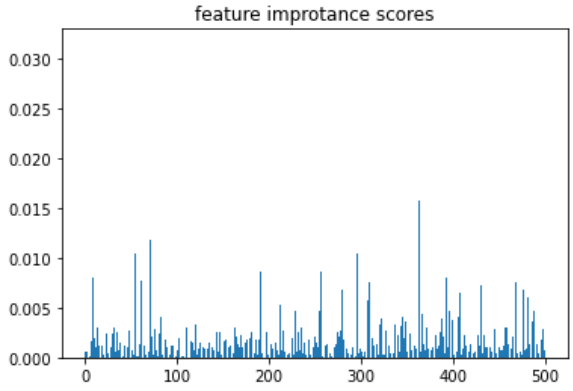

Figure 8: Feature importance scores

scores of most attributes are below 0.001. There are dozens of data with importance scores over 0.005. Considering that the random forest model has a certain degree of randomness, the picture cannot clearly indicate the true importance score of each attribute, but it still has great guiding significance.

Next, the study will use the SHAP method to test the model. The SHAP method is a method of interpreting machine learning based on Shapley values [19]. In short, the method gives the contribution of each predictor (attribute) through the Shapley value. The Figure 9 shows the twenty features with the highest contribution obtained by the SHAP method. These features come from the data set processed by PCA dimensionality reduction.

It can be seen that the contribution of these attributes is above 0.02. The three highest attribute values are feature 8 and feature 124, and their contribution degrees all exceed 0.06.

Additionally, Figure 10 is a summary plot of the features with the twenty largest contributions. From the we can see that for feature 124, when its value is relatively high, it will have a positive impact on the output; when the value is low, it will have a negative impact on the output. But feature 8 is the opposite. When its value is relatively high, the predicted value will be reduced; when its value

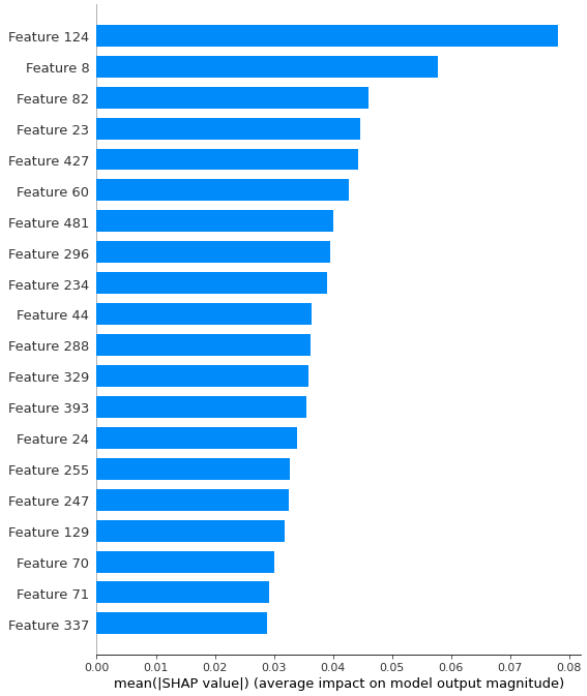

Figure 9: Feature contribution scores by SHAP

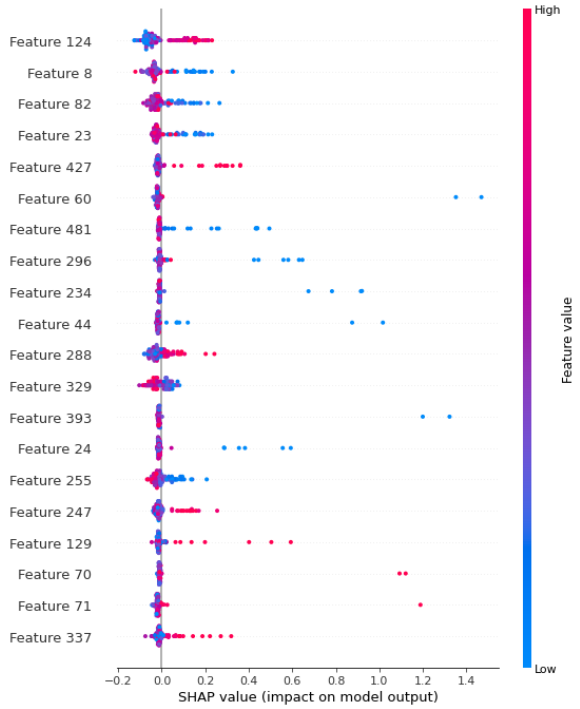

Figure 10: Summary of SHAP

is relatively small, the predicted value will be increased. Other features can also be analyzed similar to feature 124 and feature 8.

In addition, we can analyze the performance of the random forest model through the residuals. In order to make better judgments, the experiment introduces the residual plot of MLR for comparison.

Figure 11 and Figure 12 display the residual plots of the MLR and random forest model, respectively. First look at the residual plot of the MLR, which ranges from negative twenty to thirty. Its distribution is apparently random. The residual range of the residual map of the random forest model is from -3 to 13. It can be seen that compared to the range of residuals of the MLR. The scope of the random forest model is significantly reduced. However, most of the residuals of the random forest model are concentrated in the range of 0 to -3. From a comprehensive perspective, we can believe that the random forest model is better than the MLR in predicting PD.

To more accurately estimate the performance of the random forest model in predicting the progression of PD, we have introduced two evaluation indicators: MSE and SMAPE.

MSE is the mean of the sum of squares of all predicted values Y* subtracted from the actual values Y [20].

$$MSE = \frac{1}{n}\left(Y_i - \widehat{Y}_i\right)^2 \tag{5}$$

The lower the value of MSE, the closer the model prediction output is to the real output, that is, the model has better prediction ability.

In addition, we also select the SMAPE index to measure the prediction performance of the random forest model. The calculation

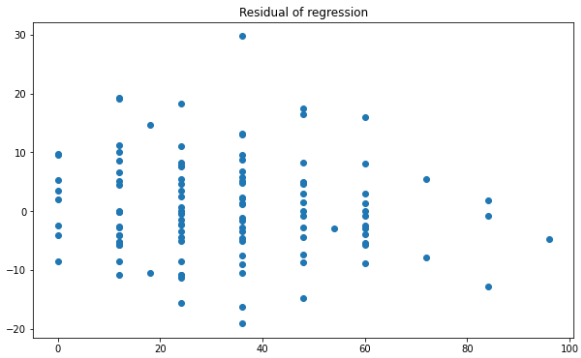

Figure 11: The Residual of regression

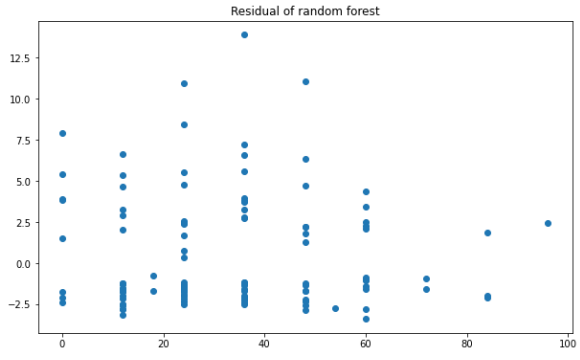

Figure 12: The Residual of random model

Table 2: Regression model performance score

MDS-UPDRS	MSE of random forests(test)	SMAPE of random forests(test)	MSE of random forests(test))	MSE of regression (test)	SMAPE of regression (test)	MSE of regression (train)
updrs_1	27.5388	69.4976	4.05227	55.8189	93.7336	6.90306
updrs_2	34.7710	95.7504	5.04956	64.0515	116.591	8.06302
updrs_3	212.621	86.5150	31.1287	395.004	105.918	50.3274
updrs_4	9.8274	152.608	1.45766	95.0316	169.427	7.62e-26

of SMAP1 is as follows:

$$SMAPE = \frac{100\backslash\%}{n} \sum_{i}^{n} \frac{\left|Y_i - \widehat{Y_i}\right|}{\left(|Y_i| + \left|\widehat{Y_i}\right|\right)/2} \tag{6}$$

SMAPE treats higher or lower prediction errors equally, so it has a certain degree of symmetry. In general, a lower value of SMAPE indicates better model performance [21].

In order to better test the performance of the random forest model, we evaluate the MSE index on the training set and the test set respectively, and evaluate the SMAPE index on the test set. Furthermore, to assess the performance of the random forest model more accurately, the experiment employed MLR as a comparative analysis.

As shown in Table 2, in addition to predicting the MSE value of updrs_4 on the test set, MLR is lower. The other index values of the random forest model are lower. For instance, when the random forest model predicts updrs_1 on the test set, its MSE and SMAPE values are 27.5388 and 69.4976 respectively, while the corresponding MSE and SMAPE of the MLR are 55.8189 and 93.7336 respectively. It is a good illustration that whether from the perspective of MSE or SMAPE, the random forest model has better generalization performance and stronger prediction ability in practice.

6 CONCLUSION

The primary objective of our experiment is to construct a random forest model utilizing the protein dataset of Parkinson's patients, followed by an analysis and diagnostic examination of the model. The experiment introduces UPDRS score, protein data and peptide data for Parkinson's patient diagnosis. These data are then combined to form the Parkinson dataset used in the experiments. In the experiment, in order to remove redundant data and noise in the data set, we utilize the PCA method to reduce the dimensionality of the original Parkinson's dataset. Ultimately obtaining a dataset containing 500 attributes.

Our efforts revolved around assessing the predictive capabilities of the random forest model and performing an in-depth analysis. We employed various methodologies, including residual analysis, SHAP analysis, and feature importance scores from the random forest model. Additionally, we introduced Multiple Linear Regression (MLR) as a benchmark for diagnostic performance evaluation. The results revealed that the random forest model outperformed MLR, as evidenced by lower Mean Squared Error (MSE) and Symmetric Mean Absolute Percentage Error (SMAPE) metrics, highlighting its superior generalization performance.

Through the protein data set of Parkinson's patients, the paper establishes a corresponding random forest model to forecast the progression of PD. The model can help identify and predict PD, which is of great help in intervening and treating patients' conditions. Improving the accuracy of diagnosing PD, early intervention and treatment of PD can significantly improve patients' quality of life and experience, and reduce their medical burden. In addition, it can also reduce the waste of medical resources and improve the utilization rate of medical resources. It has a positive impact on improving the social medical system. In addition, the experiment may also help identify relevant biomarkers which can help improve the accuracy of PD diagnosis and study the pathogenesis of PD for PD.

However, the experiment also has certain limitations. The dataset used in the experiment only contains the diagnosis data of 248 patients at different times, about 2,600. Smaller data sets may have an impact on the generalization performance of the model. In addition, there are many missing values in the dataset, which may affect the interpretability of the data and the effect of the random forest model. Of course, the experiment also uses the mean substitution method to deal with these missing values. It can ensure the effectiveness of the model to a certain extent.

Our experiment's outcome offers a preliminary work for utilizing protein data from Parkinson's patients to forecast the progression of PD. There remains substantial room for further exploration. Future work should focus on expanding the dataset by incorporating data from a more extensive pool of Parkinson's patients and exploring additional data sources beyond proteins. In addition, we will continue to explore the interpretability of the model so that medical personnel can better understand how the model works.

As we know, the field of utilizing protein data for PD prediction is challenging yet holds great promise. We encourage further research in this area, as collaboration and innovation can drive positive change in our society and healthcare system.

REFERENCES

[1] Armstrong, M. J., & Okun, M. S. 2020. Diagnosis and treatment of Parkinson disease: a review. Jama, 323(6), 548-560.
[2] Rocca, W. A. 2018. The burden of Parkinson's disease: a worldwide perspective. The Lancet Neurology, 17(11), 928-929.
[3] Bryant, M. S., Rintala, D. H., Hou, J. G., & Protas, E. J. 2015. Relationship of falls and fear of falling to activity limitations and physical inactivity in Parkinson's disease. Journal of aging and physical activity, 23(2), 187-193.
[4] Mencke, P., Boussaad, I., Romano, C. D., Kitami, T., Linster, C. L., & Krüger, R. 2021. The role of DJ-1 in cellular metabolism and pathophysiological implications for Parkinson's disease. Cells, 10(2), 347.

[5] Shen, J., Amari, N., Zack, R., Skrinak, R. T., Unger, T. L., Posavi, M., ... & Chen-Plotkin, A. S. 2022. Plasma MIA, CRP, and albumin predict cognitive decline in Parkinson's disease. Annals of Neurology, 92(2), 255-269.

[6] Postuma, R. B., & Montplaisir, J. 2009. Predicting Parkinson's disease–why, when, and how?. Parkinsonism & related disorders, 15, S105-S109.

[7] Alickovic, E., Subasi, A., & Alzheimer's Disease Neuroimaging Initiative. 2020. Automatic detection of alzheimer disease based on histogram and random forest. In CMBEBIH 2019: Proceedings of the International Conference on Medical and Biological Engineering, 16- 18 May 2019, Banja Luka, Bosnia and Herzegovina (pp. 91-96). Springer International Publishing.

[8] Dai, B., Chen, R. C., Zhu, S. Z., & Zhang, W. W. 2018, December. Using random forest algorithm for breast cancer diagnosis. In 2018 International symposium on computer, consumer and control (IS3C) (pp. 449-452). IEEE.

[9] Martinez-Martin, P., Rodriguez-Blazquez, C., Alvarez-Sanchez, M., Arakaki, T., Bergareche Yarza, A., Chade, A., ... & Goetz, C. G. 2013. Expanded and independent validation of the movement disorder society–unified Parkinson's disease rating scale (MDS-UPDRS). Journal of neurology, 260, 228-236.

[10] Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., ... & LaPelle, N. 2008. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. Movement disorders: official journal of the Movement Disorder Society, 23(15), 2129-2170.

[11] Brownlee, J. 2020. Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python. Machine Learning Mastery.

[12] Biau, G. 2012. Analysis of a random forests model. The Journal of Machine Learning Research, 13, 1063-1095.

[13] Saini, A. 2022, August 26. An introduction to Random Forest Algorithm for beginners. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/an-introduction-to-random-forest-algorithm-for-beginners/

[14] Das, R. 2010. A comparison of multiple classification methods for diagnosis of Parkinson disease. Expert Systems with Applications, 37(2), 1568-1572.

[15] Nilashi, M., Abumalloh, R. A., Minaei-Bidgoli, B., Samad, S., Yousoof Ismail, M., Alhargan, A., & Abdu Zogaan, W. 2022. Predicting parkinson's disease progression: Evaluation of ensemble methods in machine learning. Journal of healthcare engineering, 2022.

[16] Avcontentteam. 2023, June 26. PCA: What is Principal Component Analysis & How It Works? (updated 2023). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/

[17] Lee, T. H., Ullah, A., & Wang, R. 2020. Bootstrap aggregating and random forest. Macroeconomic forecasting in the era of big data: Theory and practice, 389-429.

[18] Kumar, A. 2022, December 7. Feature importance & random forest - python. Analytics Yogi. https://vitalflux.com/feature-importance-random-forest-classifier-python/

[19] Sundararajan, M., & Najmi, A. 2020, November. The many Shapley values for model explanation. In International conference on machine learning (pp. 9269-9278). PMLR.

[20] Brownlee, J. 2021, February 15. Regression metrics for machine learning. MachineLearningMastery.com. https://machinelearningmastery.com/regression-metrics-for-machine-learning/

[21] Kreinovich, V., Nguyen, H. T., & Ouncharoen, R. 2014. How to estimate forecasting quality: A system-motivated derivation of symmetric mean absolute percentage error (SMAPE) and other similar characteristics.