Analysis of factors associated with early stage Parkinson’s disease based on daily activities and sleeping behaviour disorder

F Nastitie, S Abdulllah*, and S Nurrohmah

Department of Mathematics, Universitas Indonesia, Kampus Baru UI, Depok 16424, Indonesia

*sarini@sci.ui.ac.id

Abstract. Parkinson’s Disease (PD) is a disorder in human movement coordination system that is characterized by motoric and non-motoric symptoms. At the late stage of PD, clinical diagnosis is relatively easy to detect because the symptoms are clear-cut. However, when the symptoms are often incomplete or subtle, in the initial stage, diagnosis becomes difficult and sometimes the subject still remains undiagnosed or even misdiagnosed. This study was aimed at identifying risk factors in early stage PD based on patient daily activities and rapid eye movement behaviour disorder. Daily activities were measured using the Modified Unified Parkinson’s Disease Rating Scale (MDS-UPDRS) part I and part II and sleep behaviour disorder was measured using Rapid eye movement sleeping Behaviour Disorder Screening Questionnaire (RBDSQ). Data analysis was conducted using classification trees with CART algorithm, to classify patients into early stage PD patients or healthy control patients. Missing values were handled using k-Nearest Neighbour (kNN) method. The results were satisfactory, with the classification accuracy of 86.5%, sensitivity 80%, specificity 91.57% and AUC 0.858. It is also found that tremor, dressing difficulty, speech difficulty, RBD questionnaire score, and age are important in differentiating early stage PD from healthy control.

1. Introduction

Parkinson's Disease (PD) is the second most common neurodegenerative chronic progressive disorder after Alzheimer's Dementia which causes a decrease in the quality of life of millions of people in the world [1]. This is characterized by motor and non-motor symptoms. The most common motoric symptoms in PD known as parkinsonism are bradykinesia, rest tremor, rigidity, and postural and gait impairment while non-motoric features which often occur in PD are hyposmia, rapid eye movement (REM) behaviour disorder, constipation, and depression [2].

There are many instruments to evaluate impairment and disability (stage and severity) in PD. Among these, the Modified Unified Parkinson's Disease Rating Scale (MDS-UPDRS) and the Hoehn and Yahr (HY) scale are the most commonly used [3]. Rapid Eye Movement Sleep Behaviour Disorder (RBD) is often found before neuro-degenerative disorders diagnosis such as PD. This sleep disorder is characterized by disturbing dreams, or some aggressive movements during sleeping due to vivid dreams, and the lack of normal REM-sleep muscle atonia [4]. Subjects with RBD are more likely to develop neurodegenerative diseases such as PD [5].
In clinical PD diagnosis, there is not much difficulties for the advanced stage as the symptoms are clear. On the contrary, providing an accurate diagnosis is challenging in the early stages of the disease, as the symptoms are subtle, or even incomplete [6]. Data mining, however, facilitates the opportunities to computer-aided classification and diagnosis that could reduce inevitable failibilities and inherent diagnostic variabilities in healthcare, carry out early treatment for early stage PD patients and also speed up decision-making of the disease management in the early stage of the disease.

Regarding classification problem, there are a couple of data mining methods for instance logistic regression, SVM, and decision tree. Logistic regression is good in the sense of handling linear relationship of variables, but it is not optimal to overcome the effects of the interaction between variables [7]. Some other well-known method is SVM, and this method has shown to excel in the accuracy with the help of mapping to higher dimensional spaces using appropriate kernels [8]. Nevertheless, this method is not suitable for this study because of the main purpose of this study is to identify factors related to early stage PD that are easy to interpret to help improve medical personnel [9]. Naïve Bayes classifier, on the other hand, have been shown to provide insights and high accuracy in medical diagnosis [10]. Yet, its independence assumption on the relationships among the measurements are rarely met in the real data application. Decision tree attracts for its ease of interpretation and have been showing satisfactory accuracy [11-13], yet it is known for its tendency to overfit [8, 14]. The tree structure visualization of the classification rule allows for easy interpretation, no need to employ a complex mathematical calculation, for example, as in the logistic regression and SVM. As for the tendency of overfitting, we could control and evaluate it by partitioning the data into training and testing datasets. Therefore, in this study we propose to implement a decision tree, to identify people with early stage of PD and generate their profiles so that they can be differentiated from the healthy control group.

This study used the clinical scale of daily activity parts (Parts I and II) of MDS-UPDRS. Part II is chosen as it has been shown to produce important information in explaining about the early PD [14]. In accordance, we use RBDSQ along with demographic information to classify early stage PD patients from healthy normal using medical data mining method called classification tree. The performance of the classification tree method will be seen and analysed in classifying the data of healthy control PD patients and early stage patients based on accuracy values, so that the results of this study can be used as a tool for health experts in predicting Parkinson's disease.

Missing values might result in lower performance of the model, for example, low accuracy, or even inability of the model to learn that result in relatively high error in identification of such classes in the target variable [9]. Therefore, handling missing data is one of the important procedures before building the model. Several methods for handling missing values have been proposed, such as mean imputation, median imputation, and regression imputation. Mean and median imputation are easily implementable and simple but if missing value large in number it will lead to change in shape of the distribution [15]. Regression imputation is more convenient in predicting missing values as this method use regression method based on non-missing data to replace missing values, but this method is based on linear relationship assumptions so if the data is not linearly correlated the model will be biased. On the other hand, kNN imputation can predict both discrete and continuous attributes of missing value by predicting values from similar records in the same dataset based on distance function without any assumption needed [16]. Thus, we propose to implement listwise deletion and K-Nearest Neighbour imputation.

Motivated by the above-mentioned outcomes, in this paper, we propose to implement classification tree method to differentiate early PD patients from healthy control, and further identify risk factors related to the early stage PD patients, as well as a listwise deletion and K Nearest Neighbour imputation to handle missing values in the data. The aim is to find out how the model will perform when one of the two approaches is applied to data regarding PD patients. The remainder of the paper is organized in the following parts. Section 2 briefly describes the material (data set) and methods used in our work. Section 3 details experimental results and discussion and finally this paper is summarized in Section 4.
2. Method

2.1 Data set
Data of 598 patients were obtained from Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). PPMI is an observational, multi-centre study that will collect clinical and imaging data and biologic samples from various cohorts, that can be used by scientists to establish markers of disease progression in PD [17]. This dataset was downloaded on January 29, 2020 and contains patient’s data that were observed throughout 2019 with HY scale stages 0, 1 and 2 where a scale of 0 indicates a healthy control patient and 1,2 indicates an early stage of PD. The dataset consisted of 347 healthy control patients and 251 early stage PD patients.

On each subject, measurements on disease severity (clinically) and sleeping behaviour were conducted. daily activities were measured using the Modified Unified Parkinson's Disease Rating Scale (MDS-UPDRS) consisting of a total of 13 variables non-motoric (in Part I) and 13 variables motoric (in Part II). Sleeping behaviour disorder was measured using RBDSQ. The evaluations were also measured age, race, and gender information, making a total of 30 features are used as variable predictor. Each daily activity assessment is answered based on a 5-scale of severity, from 1 representing normal condition to 5 for the severe condition.

2.2 Classification Tree
Classification tree is a part of the decision tree method in classifying, where the targeted variables are categoric. In this study, we use two class targeted variables which are early PD patients or healthy control. Let data on N patients is denoted by $D = \{d_1, d_2, \ldots, d_N\}$, consisting $N_1$ people with early PD and $N_2$ healthy control, with $N_1 + N_2 = N$. For each subject, measurements on $p$ attributes (as listed in the data section) were conducted, that is $d_i = (x_1, x_2, \ldots, x_p), i = 1, 2, \ldots, N$. Classification tree aims to partition these $N$ subjects into several groups with certain rules, such that subjects in the same group are as homogeneous as possible, that is, consist (almost all) early PD patients, or the other way around.

![Figure 1. Example of classification tree structure.](image)

Figure 1 illustrates classification tree model which includes a single binary target variable early stage PD or healthy control.

Classification tree is a simple flowchart that selects class labels of an output variable using the values of one or more input variables. The process starts at the root nodes, internal nodes, and recursively progress until it reaches the leaf nodes with class labels [18]. Other terms to know are the parent node and child node. Parent nodes describe the nodes that make up the branch and the nodes of the branch are called child nodes [19]. At each node, a split condition is applied. In this study, we use the “goodness of split criterion” which is derived from the notion of impurity. Impurity is measured for each split and the
lowest impurity value of the split is chosen. To measure the impurity value, we used Gini index with the formula showed in Algorithm 1. If $D$ is a dataset with $m$ different class labels then Gini is defined as

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

(1)

$P_i$ denotes a relative frequency if class $i$ in $D$. If the dataset split on variable $A$ into two subsets $D_1$ and $D_2$ with sizes $N_1$ and $N_2$ respectively, gini is calculated as

$$Gini_A(D) = \frac{|N_1|}{|N|} Gini(D_1) + \frac{|N_2|}{|N|} Gini(D_2)$$

(2)

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

(3)

The splitting stops when the nodes generated by their parent nodes have minimum values of observation which have been specified at 20/3 observation at least or the class of the node is already homogeneous, that is, only one class of the target variable is present and the threshold for Gini index difference for a split is 0.01[20]. After the tree was built, we determine the class for each terminal node. The majority class will be set as the label for the terminal node.

The result of classification is summarized in a confusion matrix, where the result of the class prediction is compared with the actual class [21]. From this confusion matrix, evaluation metrics can be derived to assess the performance of the classification model, in this study, we used accuracy to evaluates model in general when predicting all true class, sensitivity evaluates the model in predicting the positive class which is early stage PD in this study, while specificity evaluates model in terms of the negative class which is health control. Table 1 shows the confusion matrix for a binary classification problem.

| Predicted Negative | Predicted Positive |
|--------------------|--------------------|
| Actual Negative    | True Negative (TN) | False Positive (FP) |
| Actual Positive    | False Negative (FN) | True Positive (TP) |

The calculation below shows how the score of accuracy, sensitivity, and specificity were obtained.

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

(4)

$$\text{Sensitivity} = \frac{TP}{\text{Total of Positive Class}}$$

(5)

$$\text{Specificity} = \frac{TP}{\text{Total of Negative Class}}$$

(6)

2.3 K Nearest Neighbour Imputation

In this section, we briefly discuss the use of k Nearest Neighbour method to handle missing values in data. This method imputes missing value using the prediction results from k-Nearest Neighbour by using data that does not have missing value.

Let from data of size $N$, there are $m$ data points with missing values, where the missingness in each data point might differ (i.e. $x_i$ is missing for $d_{1m}$ but for $d_{2m}$ the missing attribute is $x_r$), with $l \neq r, l, r \leq p, d_{im}$ is the $i$-th missing data. Then, the procedure to impute the missing values can be
summarized in the following algorithm. In this study we used range-normalized Manhattan distance for numerical data and dice coefficient for categorical data and for the final distance of the observations we used Euclidean distance to determine the k nearest neighbours from the observation that we want to predict [20].

\[ d_m^N(i, j) = \frac{|x_{i,m} - x_{j,m}|}{\text{Range}(x_m)} \]  

(7)

\[ d_m(i, j) = \begin{cases} 
  d_m^N(i, j), & \text{scale, interval, ordinal data types in variable } m \\
  d_m^C(i, j), & \text{nominal data types in variable } m 
\end{cases} \]  

(8)

\[ d(x, y) = \sqrt{\sum_{m=1}^{p} d_m(i, j)^2} \]  

(9)

Step 1. Given the value of \( k \) which determines the number of neighbours to be used

Step 2. Based on the specific distance function, determine the k nearest neighbours from the observation you want to predict

Step 3. If the problem is a classification problem, then the target variable prediction given is the majority value of the target variable from the neighbours it has. If the problem is a regression problem, the prediction of the target variable given is the average of the target variables of the neighbours it has

In this study, we use five nearest neighbours \((k = 5)\), based on the specific distance function [16]. In this study, the variables that we want to impute is score of RBDSQ questionnaire which means the problem is a regression problem (numeric), the prediction of the target variable given is the average of the target variables of the observation neighbours.

2.4 Study Design

The process of the study will be illustrated in the flowchart in Figure 2. It shows the study design of the data analysis that starts of with collecting PD patients dataset, followed with data pre-processing which include data cleaning, transformation, and handling missing values with two different scheme (listwise deletion and kNN imputation), next we split the dataset into data training and data testing, and then we build the models on training dataset and implement the models in the testing data to evaluate the performance of the models, lastly we interpret the output of classification tree.

![Flowchart](image)

Figure 2. Flowchart of the study process.

3. Result and Discussion

Descriptive statistics were applied for demographic information, RBDSQ score, and daily activities disorder are presented in Table 2.
Table 2. Demographic and clinical features of the study

| Demography                      | %   | Mean (SD)     |
|---------------------------------|-----|---------------|
| Female                          | 46.3|               |
| Age at diagnosis with PD (year) | 61.27| (8.4)        |
| White race                      | 88.9|               |

Rapid Eye Movement Sleeping Behaviour Disorder

| RBDSQ Score                     | 3.872 (2.99) |

Non-motoric Assessment

| Cognitive Impairment            | 26  |
| Hallucination                   | 5.8 |
| Depress mood                    | 20.4|
| Anxious mood                    | 33.2|
| Apathy                          | 17.05|
| Features of DDS                 | 3.1 |
| Sleep Problems                  | 64.8|
| Daytime Sleepiness              | 58.7|
| Pain and other sensations       | 54.8|
| Urinary Problems                | 43.3|
| Constipation problems           | 34.6|
| Light Headedness on standing    | 28.2|
| Fatigue                         | 46.4|

Disorder in Motoric Assessment

| Speech problem                  | 75.8|
| Saliva and drooling             | 75.8|
| Chewing and swallowing          | 85.9|
| Eating tasks                    | 79.1|
| Dressing difficulty             | 72.3|
| Hygiene                         | 82.5|
| Handwriting                     | 66.8|
| Doing hobbies & other activities| 72.1|
| Turning in bed                  | 69.9|
| Tremor                          | 63.04|
| Getting up of bed/chair         | 65.7|
| Walking and balance             | 69.06|
| Freezing                        | 87.9|

One of the problems in this analysis is missing values that were found when assessment of rapid eye movement sleeping behaviour through RBDSQ was conducted. We found there are at least 150 observations (approximately 25%) with missing values in RBDSQ assessment from 598 of total data. In this section, data analysis of handling 150 missing values in RBDSQ evaluations using listwise deletion and kNN imputation was computed in R.

We build the classification tree model based on complete data resulting from the listwise deletion and kNN method. We prefer the results from kNN imputation than listwise deletion, as the evaluation metrics are more satisfactory, as presented in Table 3. Thus, the interpretation will be based on this result.

Table 3. Classification trees performance using two schemes on handling missing values in data test

|                      | Listwise deletion | kNN Imputation (k=5) |
|----------------------|-------------------|----------------------|
| Accuracy             | 82%               | 86.5%                |
| Sensitivity          | 84%               | 80%                  |
| Specificity          | 80%               | 91%                  |
|                      |                   |                      |
Table 1 shows the classification tree models using two different scheme of handling missing values in terms of accuracy, sensitivity, and specificity. KNN imputation in general gives a better performance of the model rather than listwise deletion on handling 25% missing values of RBDSQ score. Accuracy of the classification model, 82% shows that the model is good enough to classify early PD patients and healthy control. In terms of predicting the positive class (early PD patients), the sensitivity shows out of all PD patients, the model is correctly predicted 80% of the early PD patients while specificity evaluates the model which correctly predicted 91% in terms of predicting healthy control patients. Figure 3 shows the output classification tree model.

Figure 3. Output of classification tree model.

Based on the above figure, 5 out of 30 features were selected as the most important ones to explain the incidence of early PD patients. Dressing difficulty (NP2DRES) levels was selected as the most important feature since NP2DRES levels have a positive correlation with early stage PD patients. PD patients who are likely to be classified as early PD are the ones who have dressing difficulty even in the lowest severity (terminal node 1), or patients who have tremor disturbance and score of RBDSQ ≥ 3 (terminal node 2), or patients who have speech difficulty and age < 70 (terminal node 4). Note that the terminal node 1 produced the purest result, as the probability of the node is normal patients only 0.06, as well as for terminal node 5 shows that the probability of the node is normal is 1.00. These might be the most important ones and need to be recommended to check.

4. Conclusion
We develop a classification tree for identifying people with early stage of PD and differentiate them from healthy control. Missing values were handled using listwise deletion kNN imputation, with the latter provided a better model performance and thus used for further analysis. High accuracy, sensitivity and specificity were obtained showing the goodness of the model. Risk factors for early PD were identified: dressing difficulty, resting tremor, speech difficulty, rapid eye movement sleeping behaviour disorder, age and turning over in bed difficulty. We further used these factors to generate profiles for people with early PD.
Acknowledgements
This research was funded by Directorate of Research and Development of Universitas Indonesia (DRPM UI) as a grant of Publikasi Terindeks Internasional (PUTI) Prosiding 2020 No. NKB-977/UN2.RST/HKP.05.00/2020. Authors wishing to acknowledge assistance or encouragement from colleagues, special work by technical staff, and financial support from the Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Indonesia. Authors also wishing to acknowledge PPMI, a landmark observational clinical study of PD which is funded by Michael J. Fox Foundation for Parkinson’s Research for providing access for the data.

References
[1] Nussbaum R L and Ellis C E 2003 Alzheimer’s disease and Parkinson's disease The New England Journal of Medicine 348 1356-1364
[2] Massano J and Bhatia K P 2012 Clinical approach to Parkinson’s disease: features, diagnosis, and principles of management Cold Spring Harbor perspectives in medicine 2 008870
[3] Postuma R B, Gagnon J F, Vendette M, Fantini M L, Massicotte-Marquez J and Montplaisir J 2009 Quantifying the risk of neurodegenerative disease in idiopathic REM sleep behavior disorder Neurology 72 296-1300
[4] Booij J, De Jong J, De Bruin K, Knol R, De Win M M, and van Eek-Smit B L 2007 Quantification of striatal dopamine transporters with 123I-FP-CIT SPECT is influenced by the selective serotonin reuptake inhibitor paroxetine: a double-blind, placebo-controlled, crossover study in healthy control subjects Journal of Nuclear Medicine 48 359-366
[5] Sixel-Döring F, Liepe K, Mollenhauer B, Trautmann E, and Trenkwalder C 2011 The role of 123 I-FP-CIT-SPECT in the differential diagnosis of Parkinson and tremor syndromes: a critical assessment of 125 cases Journal of Neurology 258 2147-2154
[6] Choudhury A and Kosorok M R 2020 Missing Data Imputation for Classification Problems (preprint arXiv:2002.10709)
[7] Otero F E, Freitas A A and Johnson C G 2012 Inducing decision trees with an ant colony optimization algorithm Applied Soft Computing 12 615-3626
[8] Sperandei S 2014 Understanding logistic regression analysis Biochimia medica 24 12-18
[9] Yuan X, Yuan X, Yang F, Peng J, and Buckles B P 2003 Gene Expression Classification: Decision Trees vs. SVMs FLAIRS conference 2003 12-14 May (Florida, USA) 92-97
[10] Faiziyah N A, Abdullah S, and Nurrohmah S 2020 Reviewing the consistency of the Naïve Bayes Classifier’s performance in medical diagnosis and prognosis problems AIP Conference Proceedings 2242 030019
[11] Christianti D, Abdullah S, and Nurrohmah S 2019 Bayes Risk Post-Pruning in Decision Tree to Overcome Overfitting Problem on Customer Churn Classification Proceedings of the 1st International Conference on Statistics and Analytics, ICSA 2019 2-3 August (Bogor, Indonesia)
[12] Nurrohman A, Abdullah S, and Murfi H 2020 Parkinson’s disease subtype classification: Application of decision tree, logistic regression and logit leaf model AIP Conference Proceedings 2242 030015
[13] Abdullah S, White N, McGree J, Mengersen K, and Kerr G 2019 Assessing the predictive ability of the UPDRS for falls classification in early stage Parkinson’s disease (preprint arXiv:1910.01313)
[14] Abdullah S and Prasetyo G V 2020 Easy Ensemble with Random Forest to Handle Imbalanced Data in Classification Journal of Fundamental Mathematics and Applications 3 39-46
[15] Jadhav A, Pramod D and Ramanathan K 2019 Comparison of performance of data imputation methods for numeric dataset Applied Artificial Intelligence 33 913-933
[16] Batista G E and Monard M C 2002 A Study of K-Nearest Neighbour as an Imputation Method His 87 251-260
[17] Marek K, Jennings D, Lasch S, Siderowf A, Tanner C, …, and Poewe W 2011 The parkinson progression marker initiative (PPMI) Progress in neurobiology 95 629-635
[18] James G, Witten D, Hastie T, and Tibshirani R 2013 *An introduction to statistical learning* (Springer, New York)

[19] Tangirala S 2020 Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm International *Journal of Advanced Computer Science and Applications* 11

[20] Therneau T, Atkinson B, and Ripley B 2019 *Rpart: Recursive partitioning for classification, regression and survival trees R package version 4.1-15*

[21] Rokach L and Maimon O 2005 *Decision trees. In Data mining and knowledge discovery handbook* (Springer: Boston MA)