Identification of risk factors for impulse-control disorder symptoms in patients with Parkinson’s disease

D A Setianingsih, S Abdullah*
Department of Mathematics, Faculty of Mathematics and Natural Sciences, University Indonesia, Kampus Baru UI, Depok 16424, Indonesia

Abstract. Parkinson's disease (PD) is one of the motoric neurodegenerative disorders that has the fastest-growing prevalence, disability, and death rate compared to other neurological disorders. Globally, from 1990 to 2015, the number of individuals with PD increased by 118%, up to around 6.2 million. There are several motor symptoms and non-motor symptoms that occur in Parkinson’s disease. One of the non-motor symptoms that occur is Impulse-Control Disorder (ICD). In PD, there are 4 main symptoms of ICD that often occur such as pathological gambling, binge-eating, compulsive buying, and compulsive sexual behaviour. ICD is often discovered when the treatment of PD begins, so this research focused on the incidence of ICD as a result of the treatment of PD. The purpose of this study is to identify risk factors for the type and number of ICD symptoms that occur in patients with PD. Both binary and multiclass classification tree methods were implemented to identify risk factors associated with the incidence of ICD for PD patients. Imbalance data problems that arose were handled with Synthetic Minority Over-sampling Technique (SMOTE). The results obtained show that the STAI-Trait total score is a risk factor that always appears for each type of, and the number of type of, ICD symptoms that appear. Moreover, risk factors that only appear in several symptoms are the length of education taken, STAI-State total score, age, duration of PD, SCOPA-AUT total score, MOCA total score, and MDS-UPDRS 3 total score. While, risk factors that only appear for a particular symptom are family history of PD, DAT binding ratio, and dopamine agonist treatment.

1. Introduction
Parkinson's disease is a neurodegenerative disorder that occurs due to damage or death of nerve cells that contain dopamine in the part of the brain that coordinates motoric systems. According to the Global Burden of Disease, Injuries, and Risk Factors Study in 2015, Parkinson's disease is one of the neurological disorders that has the fastest-growing incidence, disability, and mortality rate compared to others. The number of people with Parkinson's disease by 2015 increased globally by 118% compared to 1990 and was recorded up to 6.2 million [1]. PD characterized by motor and non-motor symptoms. Motor symptoms such as stiffness of muscles and joints (rigidity), tremor at rest, postural instability, and slowness of movement (bradykinesia) are considered as main symptoms of PD [2].

Apart from experiencing motor symptoms, PD patients also experience non-motor symptoms. One of the non-motor symptoms that occur in patients with PD is Impulse Control Disorder (ICD). American Psychiatric Associations, in the Diagnostic and Statistical Manual of Mental Disorder IV (DSM-IV), describes ICD as a behaviour characterized by the failure to control oneself from the spontaneous urge to do something repetitive and compulsive that can harm oneself or others. There are 4 main symptoms of ICD that often occur in PD, namely pathological gambling, binge-eating, compulsive buying, and compulsive sexual behaviour [3].
ICD is one of the non-motor symptoms that is often underestimated because the incidence of ICD has been identified in a minority of PD patients [4]. The incidence of ICD is difficult to identify because many patients may not reveal the existence of an ICD to care providers for reasons of shame, denial, and motivation to continue the behaviour or other reasons. Hence, the true prevalence might be higher than existing estimates. A large cross-sectional multicenter study of 3090 patients with PD shows that at least 1 ICD symptom was identified in 13.6% patients across the United States and Canada [5].

Managing of ICD is quite difficult because patients often conflict with their urges, to do or to stop their behaviour [6]. Therefore, the best first approach to managing ICD is prevention, one of which is by informing patients and their families about the risk factors for ICD in PD. The associated risk factors for the incidence of ICD in patients with PD have also been identified by previous studies, such as unmarried status, younger age, active smokers, and a history of smoking, violence, or gambling were identified as risk factors [7]. Other studies have identified REM sleep behaviour disorder [8], anxiety [9], and depression [10] as risk factors associated with the incidence of ICD.

The incidence of ICD in PD is often found when treatment had started. Several successful studies have found that any form of Dopamine Replacement Therapy (DRT) is associated with the development of ICD, especially Dopamine Agonist (DA) [11]. Even treatment with DA was 2 to 3.5 times more likely to have an ICD than treatment with levodopa [5]. Therefore, this study focuses on the incidence of ICD as a treatment effect in PD patients.

The application of data mining to health data can provide support information for the best clinical decisions made by medical experts. With the use of data mining methods, the support information can be found based on useful pattern of features from the data studied [12]. A very important issues is how to classify large volumes of health data. A classification model provides support for identification of reliable relations between features and outcome based on the similarities that present in data [13]. An accurate model can inform both patients and medical experts about the identified risk of certain diseases in order that can be used as a guide in making decisions, especially regarding treatment. Several classification methods have been shown to produce such high accuracy in classification, such as logistic regression [14, 15, 16], support vector machines (SVM) [17], naïve Bayes classifier [18], decision tree [19, 20], and random forest [21, 22]. One of the popular classification methods among researchers is the decision tree, (for classification problem usually called by classification tree). This method can provide accurate results and easily understood by people who are unfamiliar with statistics and machine learning. Classification tree can be easily interpreted even for multiclass cases. Moreover, in this method there is no need to employ a complex mathematical calculation, for example, as in the logistic regression method. Therefore, since the results of this study to inform patients and their families as well as medical experts, the classification tree method are chosen as the data mining technique used in this study.

Motivated by the above-mentioned outcomes, this study aims to identify the risk factors for each type and the number of ICD symptoms that appear in patients with PD who have received treatment using both binary and multiclass classification trees method. To overcome the imbalanced data, we propose on implementing the Synthetic Minority Over-sampling Technique (SMOTE) technique, which shown to successfully handle this problem in classification [23]. The remainder of the paper is organized in the following parts. Section 2 briefly describes the methods used in this study. Section 3 details material (data set), experimental results, and discussion. Finally, this paper is summarized in Section 4.

2. Method
2.1. Classification Tree
Decision tree is a supervised learning method that classifies a population into branch-like segments by a series of rules or questions called the splitting rule. This method could be conveniently visualized as an inverted tree with a root node, internal nodes, and terminal nodes that easy to interpret. Since the objective of this study is to predict whether a PD patient is having any type of ICD symptoms, we will use the classification tree. An example of classification tree and its structure is shown in Figure 1.
The splitting rule in classification tree are chosen by choosing a point in the features to be the splitting point that would split parent nodes into purer child nodes (more homogenous class) [24]. In this study, the feature as the splitting variable are chosen based on the splitting rule called Gini index criterion. A feature that produces the highest Gini index difference between the parent and child nodes is sought. Stopping criteria must be applied to prevent the model from becoming overly complex. Achieving class homogeneity, the minimum number of observations for the terminal node is 20/3, and the threshold for Gini index difference for a split is 0.01 are used as stopping criteria in this study. After the tree was built, class prediction on terminal node can be made based on the most frequent class.

Stopping criteria must be applied to prevent the model from becoming overly complex. A achieving class homogeneity, the minimum number of observations for the terminal node is 20/3, and the threshold for Gini index difference for a split is 0.01 are used as stopping criteria in this study. After the tree was built, class prediction on terminal node can be made based on the most frequent class.

The performance of the classification model will be evaluated based on the accuracy, sensitivity, and specificity score that can be obtained by calculation the confusion matrix [25]. Accuracy is used to measure the ability of classification model to predicting each of class correctly in general, sensitivity shows the ability of classification model to predicting the positive class correctly which is PD patients without any ICD symptom in this study, and specificity shows the ability of classification model to predicting the negative class correctly which is PD patients with any ICD symptoms.

Table 1. Confusion matrix

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

Table 1 shows the confusion matrix for a binary classification problem. For multiclass problem will be projected as binary problem that showed one class vs all. The calculation below shows how the score of accuracy, sensitivity, and specificity are obtained.

\[
\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP}
\]  \hspace{1cm} (1)

\[
\text{Sensitivity} = \frac{TP}{TP + FN}
\]  \hspace{1cm} (2)

\[
\text{Specificity} = \frac{TN}{TN + FP}
\]  \hspace{1cm} (3)

2.2. Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE is an over-sampling approach in which the minority class is over-sampled by creating synthetic samples based on the principle of k-nearest neighbor, so the data class is balanced [26]. For synthetic sample features that are numeric, it will be generated by calculating the difference between the feature
vector under consideration and one of the nearest k-neighbors selected as the reference point. Then, multiply the difference by a random number between 0 and 1 and then add it to the main observation, you will get a new synthetic sample. This process can be written in the form of an equation

\[ x_{\text{new}} = (\hat{x}_i - x_i) \times \delta + x_i, \]

(4)

where \( x_i \) is the feature vector under consideration, \( \hat{x}_i \) one of the nearest k-neighbors of \( x_i \) was selected as the reference point, and \( \delta \in [0,1] \) is a random number. Meanwhile, synthetic sample features that are categorical will be selected based on the most voted among the nearest k-neighbors. In this study, we use \( k = 5 \) nearest neighbor and the selection of parameters (\( \alpha \) for over-sampling the minority class and \( \gamma \) for under-sampling the majority class) is adjusted so that the number of observations in the minority data class approaches the number of observations in the majority data class. An illustration of how SMOTE works is shown in Figure 2.

Figure 2. Illustration of how smote method works
Source: [27]

2.3. Study Design
The data is processed using R 3.6.3 [28] implementing caret [29] package for building the model and the DMwR [30] package for rebalancing data SMOTE. The process of the study will be illustrated in the flowchart in Figure 3.

Figure 3. Flowchart of study design

3. Result and Discussion
3.1. Data
The data used in this study were obtained from the Parkinson’s Progression Markers Initiative Markers (PPMI) database per-15 February 2020. The data can be access through www.ppmi-info.org/data. PPMI is an observational clinical study designed to establish a comprehensive set of clinical, imaging and biosample data that will be used to define biomarkers of PD progression. PPMI is taking place at 34 clinical sites across the United States, Europe, Israel, and Australia.

The total data used were 230 observations of patients with Parkinson’s disease who had received treatment and had a DAT-Scan medical record. Patients who get chosen as observations are also limited to patients who performed motor tests after receiving Parkinson’s disease treatment. This research data has 22 variables, which consist of 4 target variables, namely the type of ICD symptoms (compulsive sexual behaviour, compulsive buying, and binge-eating) and the number of ICD symptoms that appear in PD patients. The proportion of data classes on each target variable is shown in Figure 4.
The other 18 variables are explanatory variables, which include patient demographic information set, motor, non-motor, NHY rating scale, PD treatment, and DAT scanning results. Patient demographic information includes gender, age, length of education (years), duration of PD (years), and family history of having PD. The severity of motor symptoms is measured by the total MDS-UPDRS score 3. Non-motor information consists of non-motor symptoms that appear in patients with PD, such as depression was assessed using the Geriatric Depression Scale (GDS-15), the severity of anxiety was assessed by the total score of the State-Trait Anxiety Inventory (STAI) as a trait and state, the severity of autonomic dysfunction was assessed by the total score of the Scale for Outcomes in Parkinson's Disease-Autonomic dysfunction (SCOPA-AUT), the severity of cognitive impairment was assessed by the total score of the Montreal Cognitive Assessment (MoCA), rapid eye movement sleep behaviour disorder (RBD) was assessed by the RBD-Screening Questionnaire (RBDSQ), and excessive daytime sleepiness was assessed by the Epworth Sleepiness Scale. The NHY rating scale information measures the staging of PD. PD treatment information consists of treatment with Levodopa, Dopamine Agonist, and others. The DAT scan results measure the DAT binding ratio in the brain’s striatum.

3.2. Rebalancing Data Strategy
The data class of each target variable in this study is not balanced as seen in Figure 4. The classification model will not give good performance if there are imbalance data problems, especially the minority class is the focus of the problem being studied [23]. Therefore, a data rebalancing strategy will be applied to overcome this problem with SMOTE method. The performance of the classification tree before and after rebalancing strategy using SMOTE is shown in Table 2.

| Target Variable                   | Before Rebalancing SMOTE | After Rebalancing SMOTE |
|-----------------------------------|--------------------------|-------------------------|
| Binge-Eating (BE)                 |                          |                         |
| Accuracy                          | 0.8957                   | 0.8979                  |
| Sensitivity                       | 1                        | 0.8927                  |
| Specificity                       | 0                        | 0.9028                  |
| Compulsive Buying (CB)            |                          |                         |
| Accuracy                          | 0.9609                   | 0.8673                  |
| Sensitivity                       | 1                        | 0.8507                  |
| Specificity                       | 0                        | 0.8843                  |
| Compulsive Sexual Behaviour (CSB) |                          |                         |
| Accuracy                          | 0.9391                   | 0.8847                  |
| Sensitivity                       | 1                        | 0.8186                  |
| Specificity                       | 0                        | 0.9524                  |
| The Number of ICD Symptoms        |                          |                         |
| No ICD                            |                          |                         |
| Accuracy                          | 0.7739                   | 0.6440                  |
| Sensitivity                       | 0.9110                   | 0.8903                  |
| Specificity                       | 0.3077                   | 0.8220                  |
| 1 Symptom                         |                          |                         |
| Accuracy                          | 0                        | 0.5714                  |
| Sensitivity                       | 1                        | 0.8879                  |
| Specificity                       |                          | 0.8743                  |
| >1 Symptoms                       |                          |                         |
| Accuracy                          | 0.7282                   |                         |
| Sensitivity                       | 0.6667                   |                         |
| Specificity                       | 0.8220                   |                         |

Pay close attention to Table 2. The accuracy and sensitivity values in several classification models before the data rebalancing strategy applied were greater than after rebalancing. However, the classification models showed the large differences between the sensitivity and specificity values. The data class of each type of ICD symptoms even shows the specificity 0%. This means that the classification models before rebalancing were not able to predict the negative classes which is the PD patients with any ICD symptom. Also, for the number of ICD symptoms, the sensitivity 0% was obtained in the data class "1 symptom", which indicates that the classification tree model before rebalancing was
not able to predict the class "1 symptom" correctly at all. Therefore, the rebalancing strategy SMOTE in the classification tree model is suggested to use for predicting the data classes.

3.3. Identify Risk Factor for The Type and Number of ICD Symptoms

The application of the binary classification tree method using SMOTE in predicting the symptom binge-eating is shown in Figure 5.

![Figure 5. Classification tree results for symptom binge-eating (BE).](image)

Based on the tree in the Figure 5 above, seven important features were found as risk factors associated with symptom binge-eating. There are the total STAI-Trait score, age, the length of education, DAT binding ratio in the brain's striatum, the total MOCA score, duration of PD, and the total SCOPA-AUT score. Patient with PD who are likely to predicted have symptom BE are the ones who have total STAI-Trait score in range 31 - 56, age < 74, the length of education < 13, DAT binding ratio in the brain’s striatum ≥ 0.82, and the duration of PD ≥ 2.9, OR, patients who have total STAI-Trait score in range 31 - 56, age in range 67 – 74 years old, the length of education ≥ 13, total MOCA score < 30, and total SCOPA-AUT score ≥ 14, OR, patients who have total STAI-Trait score in range 31 - 56, age < 67, the length of education ≥ 13, and total MOCA score < 30. Pay attention to terminal node 10. Terminal node 10 produces the purest result, as 90% of the node is predicted have symptom BE, so the rules for this node might be the most important ones and recommended to be checked on.

![Figure 6. Classification tree results for symptom compulsive buying (CB).](image)  ![Figure 7. Classification tree results for symptom compulsive sexual behaviour (CSB).](image)

Figure 6 shows the tree results for symptom compulsive buying. Several important features were found as risk factors associated with symptom compulsive buying. There are the total STAI-Trait score, family history of PD, age, and the total STAI-State score. The tree results for symptom compulsive sexual behavior are shown in Figure 7. There are several important features were found as risk factors associated with symptom compulsive sexual behavior, such as the total SCOPA-AUT score, the length of education, the total STAI-Trait score, the total MDS-UPDRS 3 score, the duration of PD, and the total STAI-State score.
The application of the multiclass classification tree method using SMOTE in predicting the number of ICD symptoms is shown in Figure 8.

Based on the tree in the Figure 8 above, the important features were found as risk factors associated with the number of ICD symptoms are the total STAI-Trait score, the length of education, the total STAI-State score, the total MOCA score, the total MDS-UPDRS 3 score, and dopamine agonist treatment.

4. Conclusion
In this study, classification tree methods are implemented to identifying PD patients with any ICD symptoms. Imbalance data class were handled using rebalancing strategy SMOTE. Furthermore, we found that risk factors of ICD symptoms in people with PD are different based on the types of symptoms. For the compulsive sexual behaviour symptom, six risk factors were identified: autonomic dysfunction, anxiety (trait and state), motor symptoms, and duration of Parkinson's disease. While for compulsive buying symptom, anxiety (trait and state), family history of Parkinson's disease, and age were found to be the most explaining factors. As for binge-eating symptoms, seven risk factors were identified: anxiety as a trait, age, the length of education, the DAT binding ratio in the brain's striatum, cognitive impairment, duration of PD, and autonomic dysfunction. Incorporating all symptoms in a model, we found that anxiety (trait and state), the length of education, cognitive impairment, motor symptoms, and dopamine agonist treatment were highly associated with the number of ICD symptoms that occur in people with PD.

5. Acknowledgments
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 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.

References
[1] Group G B D 2015 N D C 2017 Global, regional, and national burden of neurological disorders during 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015 Lancet. Neurol. 16 877–97
[2] Jankovic J 2008 Parkinson’s disease: Clinical features and diagnosis J. Neurol. Neurosurg. Psychiatry
[3] American Psychiatric Association 2000 Diagnostic and Statistical Manual of Mental Disorders, 4th Ed. DSM-IV-TR
8

[4] Leeman R F and Potenza M N 2011 Impulse control disorders in Parkinson’s disease: Clinical characteristics and implications Neuropsychiatry (London)

[5] Weintraub D, Koester J, Potenza M N, Siderowf A D, Stacy M, Voon V, Whetettecky J, Wunderlich G R and Lang A E 2010 Impulse control disorders in Parkinson disease: A cross-sectional study of 3090 patients Arch. Neurol

[6] Schreiber L, Odlaug B L and Grant J E 2011 Impulse control disorders: Updated review of clinical characteristics and pharmacological management Front. Psychiatry

[7] Weintraub D, Oehilberg K A, Katz I R and Stern M B 2006 Test characteristics of the 15-item Geriatric Depression Scale and Hamilton Depression Rating Scale in Parkinson disease Am. J. Geriatr. Psychiatry

[8] Fantini M L, Macedo L, Zibetti M, Sarchioto M, Vidal T, Pereira B, Marques A, Debilly B, Derost P, Ulla M, Vitello N, Cicolin A, Lopiano L and Durif F 2015 Increased risk of impulse control symptoms in Parkinson’s disease with REM sleep behaviour disorder J. Neurol. Neurosurg. Psychiatry

[9] Hurt C S, Alkufri F, Brown R G, Burn D J, Hindle J V., Landau S, Wilson K C and Samuel M 2014 Motor phenotypes, medication and mood: Further associations with impulsive behaviours in Parkinson’s disease J. Parkinsons. Dis.

[10] Voon V, Hassan K, Zuoowski M, De Souza M, Thomsen T, Fox S, Lang A E and Miyasaki J 2006 Prevalence of repetitive and reward-seeking behaviours in Parkinson disease Neurology

[11] Vela L, Martínez Castrillo J C, García Ruiz P, Gasca-Salas C, Macías Macías Y, Pérez Fernández E, Ybot I, Lopez Valdés E, Kurtis M M, Posada Rodriguez I J, Mata M, Ruiz Huete C, Eimil M, Borrué C, del Val J, López-Manzanares L, Rojo Sebastian A and Marasescu R 2016 The high prevalence of impulse control behaviours in patients with early-onset Parkinson’s disease: A cross-sectional multicenter study J. Neurol. Sci.

[12] Waljee A K, Higgins P D R and Singal A G 2014 A primer on predictive models Clin. Transl. Gastroenterol

[13] Milovic B 2012 Prediction and decision making in Health Care using Data Mining Int. J. Public Heal. Sci.

[14] Abdullah S, White N, McGree J, Mengersen K, Kerr G. Assessing the predictive ability of the UPDRS for falls classification in early stage Parkinson's disease. arXiv preprint arXiv:1910.01313. 2019 Oct 3.

[15] Abdullah S, McGree J, White N, Mengersen K, Kerr G. Factors associated with injurious from falls in people with early stage Parkinson's disease. arXiv preprint arXiv:1910.02379. 2019 Oct 6.

[16] Adha MR, Nurrohmah S, Abdullah S. Multinomial Logistic Regression and Spline Regression for Credit Risk Modelling. InJournal of Physics: Conference Series 2018 Nov (Vol. 1108, No. 1, p. 012019). IOP Publishing Ltd.

[17] Sarini S, McGree J, White N, Mengerson K, Kerr G. Comparison of decision tree, support vector machines, and Bayesian network approaches for classification of falls in Parkinson’s disease. International Journal of Applied Mathematics and Statistics. 2015;53(6):145-51.

[18] Fauziyyah NA, Abdullah S, Nurrohmah S. Reviewing the consistency of the Naïve Bayes Classifier’s performance in medical diagnosis and prognosis problems. In AIP Conference Proceedings 2020 Jun 1 (Vol. 2242, No. 1, p. 030019). AIP Publishing LLC.

[19] Nurrohman A, Abdullah S, Murfi H. Parkinson’s disease subtype classification: Application of decision tree, logistic regression and logit leaf model. In AIP Conference Proceedings 2020 Jun 1 (Vol. 2242, No. 1, p. 030015). AIP Publishing LLC.

[20] Christianti D, Abdullah S, Nurrohmah S. Bayes Risk Post-Pruning in Decision Tree to Overcome Overfitting Problem on Customer Churn Classification.

[21] Dewi KC, Murfi H, Abdullah S. Analysis Accuracy of Random Forest Model for Big Data—A Case Study of Claim Severity Prediction in Car Insurance. In2019 5th International Conference on Science in Information Technology (ICSITech) 2019 Oct 23 (pp. 60-65). IEEE.
[22] Latifah E, Abdullah S, Soemartojo SM. Identifying of factor associated with parkinson's disease subtypes using random forest. In Journal of Physics: Conference Series 2018 Nov (Vol. 1108, No. 1, p. 012064). IOP Publishing Ltd.

[23] Abdullah S and Prasetyo G 2020 Easy ensemble with random forest to handle imbalanced data in classification J. Fundam. Math. Appl.

[24] Rokach L and Maimon O 2008 Data mining with decision trees: theory and applications

[25] Han J, Kamber M and Pei J 2012 Data Mining: Concepts and Techniques

[26] Chawla N V., Bowyer K W, Hall L O and Kegelmeyer W P 2002 SMOTE: Synthetic minority over-sampling technique J. Artif. Intell. Res.

[27] Ma H, Huang W, Jing Y, Yang C, Han L, Dong Y, Ye H, Shi Y, Zheng Q, Liu L and Ruan C 2019 Integrating growth and environmental parameters to discriminate powdery mildew and aphid of winter wheat using bi-temporal Landsat-8 imagery Remote Sens.

[28] R Core Team 2020 R: A language and environment for statistical computing.

[29] Kuhn M 2020 caret: Classification and Regression Training. R Package. version 6.0-86. https://CRAN.R-project.org/package=caret.

[30] Torgo L 2016 Data mining with R: Learning with case studies. Chapman and Hall/CRC. http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR.