Laterality condition analysis on non-arteritic anterior ischemic optic neuropathy patient in one of the hospital in Jakarta with medical data mining

D B Unggul¹, S Abdullah¹ and A Rachman²

¹Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Depok 16424, Indonesia
²Division of Hematology and Medical Oncology, Department of Internal Medicine, Faculty of Medicine, Universitas Indonesia, Jakarta 10430, Indonesia

Corresponding author’s email: sarini@sci.ui.ac.id

Abstract. Non-arteritic Anterior Ischemic Optic Neuropathy (NAION) is a disease caused by blood shortages in the artery that supplies the optic disc. Risk factors for NAION are hypertension, obesity, diabetes, dislipidemia, smoking, hypercoagulable state, cardiovascular disease, and stroke. NAION can result from unilateral or bilateral conditions. This study will focus on the identification of important factors that could distinguish characteristics between unilateral and bilateral patients. Random forest method is applied to obtain factors that can consistently distinguish characteristic between each laterality condition. Decision trees and the logistic regression method are added to obtain the visualization of the role of each important factors in the form of classification tree and the risk comparison of patients for experiencing a certain laterality condition by using odds ratios. The important factors based on random forest model are onset, fasting blood glucose levels, high density lipoprotein levels, age, two-hour postprandial glucose levels, and low density lipoprotein levels. Based on the odds ratio, advancing age and high density lipoprotein levels will decrease the risk of patients experiencing bilateral condition; on the other hand, the risk of bilateral condition will increase if other important factors are also increased.

Keywords: Decision tree, laterality, logistic regression, random forest

1. Introduction
Non-arteritic Anterior Ischemic Optic Neuropathy (NAION) is known as an illness caused by blood shortages in the artery that supplies the optic disc. NAION often affects patients aged 40 to 60 years and is the most common optic neuropathy disorder experienced after glaucoma in adults [1, 2]. The prevalence of NAION reaches 2.3 to 10 cases per 100,000 people in the United States and 1 case per 16,000 people in China [2, 3]. 6,000 new cases are recorded annually in the United States [2].

One of the condition that is often observed and associated with the emergence of NAION is the existence of resistance in the blood vessels of patients. This resistance can be had by patients with hypertension or hypercoagulation. History of these two conditions is then often investigated as a risk factor associated with NAION. In addition to hypertension and hypercoagulation, dyslipidemia is also thought to have an important role in NAION. Dyslipidemia is defined as a condition where the amount...
of lipid is not normal [4]. This condition includes excessive low density lipoprotein (LDL) levels and insufficient high density lipoprotein (HDL) levels. High risk of atherosclerosis in people with high LDL levels [5] is thought to have a relationship with perfusion insufficiency which can lead to NAION.

The history and status of comorbidities such as stroke, cardiovascular disease, and diabetes are also thought to be associated with NAION. Stroke ranks third on the list of most common comorbidities encountered in NAION patients [6]. The relationship between cardiovascular disease and NAION is often associated with risk factors for cardiovascular disease itself, namely lipoprotein (a) [7].

Patients with diabetes usually experience conditions where blood sugar levels are very high, or what is often referred to as a state of hyperglycemia [8]. In hyperglycemic conditions, the risk of inadequate blood flow increases [8]. These conditions have the potential to cause NAION.

The patient’s lifestyle, such as smoking habits, will also be considered as a risk factor for NAION. High levels of anti-oxidants in cigarettes will trigger morphological damage to blood vessels and stimulate the formation of thrombus in blood vessels. This can cause blood flow insufficiency and can cause NAION [9]. Based on those explanations, some risk factors such as hypertension [6], diabetes mellitus [10], obesity [10], dyslipidemia [11], smoking habits [9], cardiovascular disease [7] and stroke [6] are thought to have a role in the occurrence of NAION.

NAION can be divided into two laterality conditions, namely unilateral and bilateral [1]. Those two conditions have different effects and treatments. For example, a commonly recommended treatment in patients with unilateral conditions is aspirin treatment [2]. It has been proven to reduce the risk of NAION in the second eye [2]. The difference in treatment then makes information about the patient’s laterality condition important to know. Observing and determining the number of eyes affected by NAION is difficult. That is because the doctor must observe the optic disc that is on the back of the eyeball and observe whether there are symptoms such as swelling and obstruction of blood flow.

One solution that might be used to determine a laterality condition is to observe the special characteristics of the unilateral and bilateral conditions. If the characteristics of a patient have a very strong match with the characteristics of patients with bilateral conditions, then it can be suspected that the patient has bilateral conditions. Determination of the characteristics of each conditions will be discussed in this study. This research will therefore focus on identifying factors that can consistently be used as distinguishing characteristics between unilateral and bilateral conditions. The factors included in the attempt to understand the characteristics of unilateral and bilateral conditions are demographic factors and risk factors that are suspected to be associated with the emergence of NAION. Demographic factors include age, gender, onset, visual acuity, and chief complaints from the patient.

These characteristics can be used as initial considerations to find out what steps should be taken by medical personnel. This study will also provide a comprehensive depiction of the role of each factor in the patient's laterality condition. The sight will be obtained in the form of a classification tree and odds ratio.

2. Materials and method

This section will explain about the data and the variables used in this research and what methods are involved in the analysis.

2.1. Data description

The data used in this research is the data from a hospital in Jakarta with a data collection period from 2012 to 2017. The data contains 16 measurements consisting of patient demographics and some information about risk factors. The list of variables is shown in table 1.

2.2. Random forest

Identification of the factors that can be the main characteristic of each condition will utilize the random forest method. Random forest has two main processes in its construction: bootstrapping and random feature selection [12]. Bootstrapping is a random sampling process with replacement, whereas random
feature selection is the process of taking several explanatory variables randomly that will be used as separator candidates in each split. The random forest model will display important factors but the role of each factor cannot be interpreted easily. Therefore, decision tree and logistic regression method will be used to get a description of the role of each important factor. The decision tree will produce a classification tree in a hierarchical structure while logistic regression will provide some odds ratios that can be used to quantify and compare patient risks.

2.3 Decision tree
A decision tree has three components: a root node, internal nodes, and leaf nodes. Figure 1 illustrates the output of a decision tree.

The best separator criteria for the decision tree method in this research is the Gini index and is shown in equation 1 [13]. Equation 2 shows the Gini index of the data if variable $A$ is used as a separator. Equation 3 is the reduction of impurities produced by variable $A$. Choose the variable with the smallest Gini index or the one that reduces impurities the most.

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

(1)

$$Gini_A(D) = \left(\frac{D_1}{D}\right) Gini(D_1) + \left(\frac{D_2}{D}\right) Gini(D_2)$$

(2)

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

(3)

| No. | Variable                                | Type            | Role                      |
|-----|-----------------------------------------|-----------------|---------------------------|
| 1   | Laterality Condition                    | Categorical     | Response Variable         |
| 2   | Age                                     | Numerical       | Explanatory Variable      |
| 3   | Chief Complain                          | Categorical     | Explanatory Variable      |
| 4   | Onset                                   | Numerical       | Explanatory Variable      |
| 5   | Visual Acuity                           | Categorical     | Explanatory Variable      |
| 6   | Hypertension                            | Categorical     | Explanatory Variable      |
| 7   | Smoking Habits                          | Categorical     | Explanatory Variable      |
| 8   | Cardiovascular Disease                  | Categorical     | Explanatory Variable      |
| 9   | Gender                                  | Categorical     | Explanatory Variable      |
| 10  | Stroke                                  | Categorical     | Explanatory Variable      |
| 11  | Obesity                                 | Categorical     | Explanatory Variable      |
| 12  | High Density Lipoprotein Level (HDL)    | Numerical       | Explanatory Variable      |
| 13  | Low Density Lipoprotein Level (LDL)     | Numerical       | Explanatory Variable      |
| 14  | Fasting Blood Glucose Level (GDP)       | Numerical       | Explanatory Variable      |
| 15  | Two-Hour Postprandial Level Glucose (GD2PP) | Numerical       | Explanatory Variable      |
| 16  | Hypercoagulable State                   | Categorical     | Explanatory Variable      |
The cut-off value for each separation will consider all possible points of separation from each variable. For example, suppose there is a categorical variable $B$ with 3 categories \{B_1, B_2, B_3\}. Then, separation using variable $B$ will have 3 possible splitting. The first split will divide the data into two nodes which each contain categories \{B_1\} and \{B_2, B_3\}, respectively. The second separation will form nodes \{B_2\} and \{B_1, B_3\}. The third possibility will divide data into nodes \{B_3\} and \{B_1, B_2\}. We choose one possibility of splitting that gives us the smallest Gini index. In numeric variable, variable $C$, the separation will be carried out into conditions $C < \text{splitting point}$ and $C \geq \text{splitting point}$. The chosen splitting point is a value in variable $C$ that produces the smallest Gini index when the value is used as a separator.

2.4. Logistic regression

Let $x_i = (1, x_{1i}, x_{2i}, ..., x_{pi})$ with $i = 1, ..., n$ be the values of $p$ explanatory variables from the $i$-th observation. Note that $\Pr(Y_i = 1|x_i) = \pi(x_i)$ and $\Pr(Y_i = 0|x_i) = 1 - \pi(x_i)$. The logistic regression model can be seen in equation 4 [14]. Equation 4 can then be rewritten into equation 5. The odds ratio is obtained from the estimated value of $\hat{\beta}$’s in equation 5.

$$\pi(x_i) = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})}$$  \hspace{1cm} (4)

$$\log \left( \frac{\pi(x_i)}{1 - \pi(x_i)} \right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$  \hspace{1cm} (5)

For example, the odds ratio for variable $X_1$ can be obtained by calculating $\exp(\hat{\beta}_1)$. The odds ratio can be interpreted as the increment estimation of the probability of success at every one-unit change of the value of that variable. If the odds ratio has a value between 0 and less than 1, adding one unit of the variable’s value will reduce the risk of success. If the odds ratio is more than 1, adding one unit of the variable’s value will increase the risk of success. If the odds ratio is one, then the difference in the value of that predictor variable has no effect on the change of the risk of success. In this study, the meaning of a successful event is the occurrence of bilateral conditions.
2.5. Imbalanced data and rebalancing strategy
Besides using classification methods, this research also uses oversampling and undersampling to deal with imbalanced data. Oversampling will be focused on minority class, where observations on these class will be replicated such that their number will increase. Undersampling will reduce some observations in the majority category so that the observations in both classes has the same amount [15].

2.6. Model evaluation
The evaluation of the model will consider accuracy, sensitivity, and specificity. Suppose there is a confusion matrix below (table 2).

Accuracy, sensitivity, and specificity are formulated using equation 6 through equation 8.

\[
\text{Accuracy} = \frac{TP + TN}{P + N}
\]

\[
\text{Sensitivity} = \frac{TP}{P}
\]

\[
\text{Specificity} = \frac{TN}{N}
\]

In the context of positive and negative terms, positive classes will be represented by bilateral condition and negative classes will be represented by unilateral condition.

2.7. Data analysis scheme
In order to make the scheme of data analysis clear, flowchart of analysis is given in figure 2.

3. Results and discussion
Descriptive statistics of the laterality condition of NAION patients is shown in figure 3.

It can be seen from the figure 3 that there are 21 % of patients with bilateral condition and 79 % of patients with unilateral condition. This condition indicates that there is an imbalanced data and it is necessary to handle the problem with a rebalancing strategy. But before the rebalancing strategy is applied and a random forest model is constructed, the first step is to divide the data into training data and testing data. Training data covers 75 % of all observations; the testing data used to evaluate the model will cover the remaining 25 % of all observations. After the training-testing distribution has been completed, the next step is to implement the rebalancing strategy, which uses undersampling and oversampling methods. The number of observations after rebalancing procedures are shown in table 3.

| Table 2. Confusion matrix |
|----------------------------|
| Prediction class           |
| Actual class               |
| Positive                   | Positive | Negative | Total |
| True Positive (TP)         |          |          | P     |
| False Positive (FP)        | P'       |          |       |
| False Negative (FN)        |          |          | N     |
| True Negative (TN)         |          | N'       |       |
| Total                      | P + N    |          |       |
The random forest model construction was then carried out on two datasets: the undersampling and oversampling datasets. The accuracy value of the random forest model with undersampling is 65.4 %, with sensitivity of 80 % and specificity of 61.9 %, whereas the random forest model with oversampling has an accuracy of 88.5 %, a sensitivity of 80 %, and a specificity of 90.5 %. From this result, the random forest model with oversampling has better performance than the random forest model with undersampling. The most important variables based on random forest method are shown in table 4.
From the result shown in table 4, it can be seen that four of the six important variables are laboratory examination variables. Two other important variables are the onset and age of the patient. Then it will be seen how each of these important variables classifies patients into one laterality condition. The decision tree method will be applied using only the variables in table 4. Figure 4 presents the output of the decision tree that was formed.

Here is how the decision tree in figure 4 works in classifying patients.
1. If the patient has onset $\geq$ 26, then the patient is predicted to be bilateral.
2. If the patient has onset $< 26$ and is $< 42$ years old, then the patient is predicted to be bilateral.
3. If the patient has onset $< 26$ and aged 42 to 58 years old, then the patient is predicted to be unilateral.
4. If the patient has onset $< 26$, aged $\geq$ 58 years old, and has HDL level $< 48$, then the patient is predicted to be unilateral.
5. If the patient has onset $< 26$, aged $\geq$ 58 years old, and has HDL level $\geq 48$, then the patient is predicted to be bilateral.

**Table 4. Important variables based on random forest model**

| No. | Variable                                      |
|-----|-----------------------------------------------|
| 1   | Onset.                                        |
| 2   | Fasting Blood Glucose (GDP) Level.            |
| 3   | High Density Lipoprotein (HDL) Level.         |
| 4   | Age.                                          |
| 5   | Two-Hour Postprandial Glucose (GD2PP) Level. |
| 6   | Low Density Lipoprotein (LDL) Level.          |

**Figure 4. Output of the decision tree that has been formed which only involves important variables.**
Table 5. The combination of risk factors that determine the classification of patients into unilateral or bilateral conditions.

| Prediction of laterality condition | Risk factor’s combination                                      |
|-----------------------------------|---------------------------------------------------------------|
| Bilateral                         | 1). onset ≥ 26                                                |
|                                   | 2). onset < 26, age < 42                                      |
|                                   | 3). onset < 26, age ≥ 58, HDL ≥ 48                            |
| Unilateral                        | 1). onset < 26, 42 < age ≤ 58                                 |
|                                   | 2). onset < 26, age ≥ 58, HDL < 48                            |

Table 6. The estimated value of the parameters of the logistic regression model.

| Variable | \( \exp(\beta) \) |
|----------|---------------------|
| Age      | 0.9708              |
| Onset    | 1.0136              |
| LDL      | 1.0033              |
| HDL      | 0.9905              |
| GDP      | 1.0035              |
| GD2PP    | 1.0027              |

Table 5 gives the combination of risk factors that determine the classification of patients into one of two laterality conditions (unilateral and bilateral). These combinations are determined by the classification tree in figure 4.

Based on the classification rules above, several examples will be given regarding to the classification of patients. Suppose there is a 35-year-old patient with onset less than 20 weeks. According to the classification tree in figure 4, this patient will have bilateral condition. Meanwhile, if there is a patient with onset around 35 weeks, then this patient will be classified as having bilateral conditions. If the patients have onset < 26 weeks and aged ≥ 58 years old, then classification will be based on HDL measurement value. If their HDL level is below 48, then they will be classified as having unilateral condition.

The next step is to get the odds ratio using the binary logistic regression method. The odds ratio is used to quantify the escalation of the risk of patients experiencing bilateral conditions if there is an increase in the value of these variables.

Based on table 6, if onset, LDL level, GDP level, or GD2PP level escalate, then the risk of patients experiencing bilateral conditions will increase. Meanwhile, if the patient's HDL level or age increases, then the risk of experiencing bilateral conditions will decrease. In more detail, binary logistic regression parameters can be interpreted in several points.

1. For every one-unit increase in age, the risk of emergence of bilateral conditions will increase 0.9708 times from the original risk.
2. For every one-unit increase in onset, the risk of emergence of bilateral conditions will increase 1.0136 times from the original risk.
3. For every one-unit increase in LDL level, the risk of emergence of bilateral conditions will increase 1.0033 times from the original risk.
4. For every one-unit increase in HDL level, the risk of emergence of bilateral conditions will increase 0.9905 times from the original risk.
5. For every one-unit increase in GDP level, the risk of emergence of bilateral conditions will increase 1.0035 times from the original risk.

6. For every one-unit increase in GD2PP level, the risk of emergence of bilateral conditions will increase 1.0027 times from the original risk.

4. Conclusion

This research is focused on identifying risk factors that can be the main characteristics of each laterality condition of NAION patients. Based on the random forest model with oversampling, factors that can be considered as the main characteristics are onset, GDP level, HDL level, age, GD2PP level, and LDL level. The decision tree provides some characteristics of each laterality condition where patients who meet one of the 3 conditions, which are (1) onset > 26, (2) onset < 26 and aged < 42, and (3) onset < 26, age ≥ 58, and HDL ≥ 48, will be predicted to have bilateral conditions. Patients with two other conditions like (1) onset < 26, aged between 42 and 58, or (2) onset < 26, age ≥ 58, and HDL < 48, will be categorized as patients with unilateral conditions. In addition to decision tree, binary logistic regression provides odds ratios for each important factors. Based on odds ratio, escalation of onset, LDL, GDP, and GD2PP will all increase the risk of patients experiencing bilateral conditions. On the other hand, with increasing HDL levels or patient age, the risk of experiencing bilateral conditions will decrease.

References

[1] Gerstenblith A T and Rabinowitz M P 2012 The Wills Eye Manual 6th edition (Philadelphia: Wolters Kluwer)
[2] Berry S, Lin W V, Sadaka A and Lee A G 2017 Eye Brain. 9 23-8
[3] Xu L, Wang Y and Jonas J B 2007 Eur. J. Ophthalnom. 17 459-60
[4] Nouh F, Omar M and Younis M 2019 Asian Journal of Cardiology Research 2 1-10
[5] Tomkin G H and Owens D 2012 The Open Atherosclerosis & Thrombosis Journal 5 13-21
[6] Xiao Y Y et al. 2018 Chin. Med. J. 131 2357-9
[7] Giambene B et al. 2009 Graefes Arch. Clin. Exp. Ophtalmom. 247 693-7
[8] Mouri M I and Badireddy M 2019 Hyperglycemia (Florida: StatPearls)
[9] Kim D H, Shin G R and Choi Y J 2017 Neuroophthalmology 41 68-75
[10] Chen T et al. 2013 PLoS One 8 1-7
[11] Deramo V A et al 2003 Ophthalmology 110 1041-6
[12] Breiman L 2001 Machine Learning 45 5-32
[13] Breimen L, Friedman J H, Olsens R A and Stone C J 1984 Classification and Regression Tree (New York: Chapman and Hall)
[14] Agresti A 2013 Categorical Data Analysis 3rd edition (New Jersey: John Wiley & Sons)
[15] Zhao Y, Wong Z S Y and Tsui K L 2018 J. Healthc. Eng. 2018 6275453