Classification of thalassemia data using random forest algorithm

F R Ashari¹,a), Z Rustam¹(b), F Subroto³, and A S Semendawai²

¹Department of Mathematics, University of Indonesia, Depok, 16424, Indonesia
²Harapan Kita Children and Women’s Hospital, Jakarta, 11420, Indonesia
³University of Indonesia, Depok, 16424, Indonesia

Corresponding author: a)fiftiakah.ruspian@sci.ui.ac.id, b)rustam@ui.ac.id

Abstract. Thalassemia is a blood disorder that occurred in Southeast Asia. Thalassemia cannot be cured, but early detected thalassemia with screening process is the best way to prevent thalassemia disease. If early detection is done, patients can get the right treatment. It helps them increase their life expectancy and reduce the risk of thalassemia to the next generation. In this paper, we use thalassemia data and propose a random forest method to classify thalassemia disease well and accurately. The result concludes that the random forest algorithm can give the best accuracy, precision and recall which is 100 percent by using multiple five in range of 70 to 85 percent as the training data.

1. Introduction

Thalassemia is a blood disorder that is genetically inherited from parents to their children, and is not a contagious disease [1]. Thalassemia sufferers have reduced or impaired globin chain synthesis, it causes disruption or reduced production of hemoglobin in red blood cells [2]. The word of thalassemia is derived from two Greek words, there are “Thalassa” which means “the sea” and “Haima” which means “blood” [2]. The distribution of thalassemia is called “Thalassemia belt”. Thalassemia belt includes Mediterranean across the Middle East through Southern Asia to Southeast Asia [3]. However, migrations of people caused thalassemia genes to spread throughout the world and extend to Indonesia. In Indonesia, Thalassemia carriers are 699, or 5.8% of 12,038 people who were examined. Annually, 2,500 babies are born with thalassemia in Indonesia. In October, 2016 9,131 patients with thalassemia were detected, and many have not been recorded in the Ministry of Health Republic of Indonesia [1].

Based on the two polypeptide chains, thalassemia is divided into beta (β)-thalassemia and alpha (α)-thalassemia. In beta-thalassemia, the globin chain is affected or the abnormal hemoglobin involved is beta-globin gen. Whereas in alfa-thalassemia, alfa-globin gene that affects the globin chain. Clinically, Thalassemia can be divided into three forms, there are thalassemia major, thalassemia intermedia, and thalassemia minor or trait. Thalassemia major is indicated that the patient severe anemia and they need continuous blood transfusions throughout their lives [4]. Thalassemia intermedia, characterized by mild to moderate anemia and occasionally need blood transfusions. Whereas in thalassemia minor, usually the patients don’t need blood transfusions and they look physically fit [5].

Medical checkup is important for people who look healthy to determine whether they have thalassemia or not. Therefore, random forest is proposed in this paper to classify thalassemia data. Random forest is one of the well-known cases of machine learning algorithms and unexcelled in
accuracy for classification problems, so we expect the method can classify the thalassemia data well and accurately. Moreover, the patients can get the right treatment and reduce the risk of thalassemia to the next generation. Classification with the random forest is one of the cases of supervised learning process that use labeled data [6]. It is constructed with aims to predict the class of the new case by studying the categories of case and the labels using training and testing datasets.

Supervised learning has several methods, there are decision tree, random forest, support vector machine (SVM), naïve Bayes, etc. The previous researches have been successfully classified, analyzed, diagnosed, and predicted the dataset using supervised learning. T Nadira and Z Rustam [6] implemented SVM with features selection to classify cancer data and they obtained the accuracy is 96.42 percent on breast cancer dataset and 99.99 percent on lung cancer dataset. Ulfah Aprilliani and Zuherman Rustam [7] applied random forest method to predict osteoarthritis disease and they obtained the accuracy is 86.96 percent. Mia Huljanah et al. [8] predicted prostate cancer and obtained the accuracy is 99.37 percent using random forest and 100 percent using random forest with feature selection. Zuherman Rustam and Glori [9] applied random forest method and obtained the accuracy is 100 percent. Zuherman Rustam et al. [10] analyzed gene expression data in chronic kidney disease, they obtained the accuracy 83.4 percent using SVM method, and random forest method with accuracy 7 percent. V. Panca and Zuherman Rustam [11] classified brain cancer using application of machine learning and the reported accuracy was 100 percent. Rakshita et al. [12] obtained the accuracy 94 percent, 93 percent, and 97 percent by using SVM, decision tree, and random forests, respectively. Moreover, several researches have been used on the other machine learning methods to classified, predicted, or analyzed the thalassemia data [4], breast cancer data [13], high dimensional breast cancer data [14], insurance companies [15], and cancer data [16].

2. Materials and methods

2.1 Dataset

In this paper, the dataset was obtained from Harapan Kita Children and Women’s Hospital, Indonesia. The dataset consists of 150 data patients, 82 patients are thalassemia and 68 patients are non-thalassemia. The dataset of thalassemia represented by 10 variables and target variables or class (thalassemia and non-thalassemia). The variables of thalassemia dataset are shown in Table 1.

| Variables Name      | Variables Unit   |
|---------------------|------------------|
| Hemoglobin          | g/dL             |
| Haematocrit         | Percent (%)      |
| Leukocytes          | 10^3/μL          |
| Basophils           | Percent (%)      |
| Eosinophils         | Percent (%)      |
| Rod Neutrophils     | Percent (%)      |
| Segment Neutrophils | Percent (%)      |
| Lymphocytes         | Percent (%)      |
| Monocytes           | Percent (%)      |
| Platelets           | 10^3/μL          |

2.2 Random Forest (RF)

Random Forest has some advantages, one of them is unexcelled in accuracy, can handle missing value in data and estimates the variables are important in the classification [17,18]. Random forest is a method that aims to predict the response of an observation by combine the prediction results from several decision trees. In classification problems, the prediction result of random forest based on majority votes
that is the class often arises as a result of prediction from the several trees [18]. In RF, each node of decision tree is selected randomly from the subset of variables and use those variables as candidates to find the best split for the node [19]. Figure 1 show general illustration of random forest.

Figure 1. The General Illustration of Random Forest

The random forest algorithm is constructed by several decision trees, in which each tree is grown using a bootstrap sample of the data [19]. The construction of random forest is illustrated in Algorithm 1 [12].

Algorithm 1 Random Forest (RF)

| Line | Description |
|------|-------------|
| 1.   | function RF(S, A) |
| 2.   | F ← ∅ |
| 3.   | for i ∈ 1, 2, ..., R do |
| 4.   | S(i) ← A bootstrap sample from S |
| 5.   | f_i ← RandomizedTreeLearn(S(i), A) |
| 6.   | F ← F U {f_i} |
| 7.   | end for |
| 8.   | return F |
| 9.   | end function |
| 10.  | function RandomizedTreeLearn(S, A) |
| 11.  | At each node: |
| 12.  | a ← subset of A |
| 13.  | Split best attributes in a |
| 14.  | return the learned tree |
| 15.  | end function |

After several decision trees were built, the data not in the bootstrap sample can be used as test set for that decision tree. Moreover, the result of majority voting can be predicted from several decision trees. Figure 2 present bootstrapped sample and out-of-bag sample in random forest algorithm.
In addition, OOB can be used to estimate importance of variables [19]. Variable importance can be measured by Gini index [20]. If a node \( d \) and probabilities of estimated classes is \( p(n|d) \) for \( n = 1,2,...,N \) and \( N \) is the number of classes, The Gini index is denoted as:

\[
Gini(d) = 1 - \sum_{n=1}^{N} p^2(n|d)
\]  

(1)

The Gini index value will select variables for split as the node and to know the significance of variables [7,20].

### 2.3 Confusion Matrix

In classification, the confusion matrix can be used to evaluate performance of the method. Confusion matrix show the number of samples that were correctly and incorrectly diagnosed from classification model compared to the actual results in the data [13,15]. Table 2 show the confusion matrix.

| Class       | Actual |     |
|-------------|--------|-----|
|             | Positive | Negative |
| Diagnosed   | TP | FP |
|             | FN | TN |

Based on Table 2, there are four conditions for measuring performance. True Positive (TP), it means the number of samples having thalassemia disease which are correctly diagnosed. False Negative (FN), it means the number of samples having thalassemia disease which are incorrectly diagnosed. False Positive (FP), it means the number of non-thalassemia samples which are incorrectly diagnosed. True Negative (TN), it means the number of non-thalassemia samples which are correctly diagnosed.

According to the value in confusion matrix table, we can calculate the value of accuracy, precision, and recall. Accuracy indicates that the system can classify data correctly. Precision value indicates an accurate model to predicted positive. Whereas, recall is the ratio of actual positive cases that predicted incorrectly as negative. The formula of accuracy, precision, and recall is shown in Table 3.
Table 3. The formula to evaluate performance of the method

| Parameters | Formula |
|------------|---------|
| Accuracy   | \( \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \) |
| Recall     | \( \frac{TP}{TP + FN} \times 100\% \) |
| Precision  | \( \frac{TP}{TP + FP} \times 100\% \) |

3. Experimental results

In this paper, the data of 150 thalassemia patients were used and divided into multiple five in range 50 to 85 percent as training data. The result of the random forest algorithm using nTrees = 500 decision trees, 10 variables, and target variables or class (thalassemia and non-thalassemia) is shown in Table 4.

Table 4. The result of Random Forest algorithm

| Training Data (%) | Accuracy (%) | Recall (%) | Precision (%) |
|-------------------|--------------|------------|---------------|
| 50                | 98.31        | 100        | 96.97         |
| 55                | 98.11        | 100        | 96.67         |
| 60                | 97.92        | 100        | 96.30         |
| 65                | 97.56        | 100        | 95.65         |
| 70                | 100          | 100        | 100           |
| 75                | 100          | 100        | 100           |
| 80                | 100          | 100        | 100           |
| 85                | 100          | 100        | 100           |

According to the calculation of confusion matrix from the random forest algorithm’s result, the best accuracy, recall, and precision are 100 percent by using multiple five in range 70 to 85 percent as training data. Meanwhile, all of the percentage of training data gives the best recall value, which is 100 percent. The average of accuracy, recall, and precision from the random forest algorithm using multiple five in range 50 to 85 percent as training data are 98.99 percent, 100 percent, and 98.20 percent, respectively. Figure 3 show running time of the random forest algorithm with different percentage of training data.

Figure 3. The running time of random forest algorithm with different percentage of training data.
Based on Figure 3, by using more training data the longer time is needed to get the result of the random forest algorithm and the average of the running time is 11.13 seconds. Moreover, significance of variables from the random forest can be known, that is shown in Figure 4.

![Variables Importance](image)

**Figure 4.** Significance of the variables using training data 85%.

Based on Figure 4, the highest rank of important variables is Haematocrit. Meanwhile, the descending order of the variables are Haematocrit, Rod Neutrophils, Hemoglobin, Eosinophils, Segment Neutrophils, Leukocytes, Lymphocytes, Monocytes, Basophils, and Platelets.

### 4. Conclusion

Based on the result of this method, Random forest can classify thalassemia disease well and give the best accuracy, precision and recall which is 100%. It was obtained by using multiple five in the range of 70% to 85% as training data and the average of the running time is 11.13 seconds. Meanwhile, using training data 85%, the rank of the variables importance are Haematocrit, Rod Neutrophils, Hemoglobin, Eosinophils, Segment Neutrophils, Leukocytes, Lymphocytes, Monocytes, Basophils, and Platelets.

We believe that future research can develop a new method or modify this method for predicting or classifying the other diseases. By performing this research, we hope the result can help the medical sector to classify and detect whether the patient has thalassemia or not, so the patient can receive the right treatment to increase their life expectancy and reduce the risk of thalassemia to the next generation.

### Acknowledgement

This research supported financially by the Ministry of Research and Higher Education Republic of Indonesia with PDUPT 2019 research grant scheme, ID number 1621/UN2.R3.1/HKP.05.00/2019. The authors want to thank the head of Harapan Kita Children and Women’s Hospital who helps and allows the authors to use the data.

### References

[1] Ministry of Health Republic of Indonesia Accessed: 17 August 2019 http://www.depkes.go.id/

[2] Peters M, Heijboer H, and Giordano P C 2012 Diagnosis and management of thalassemia BMJ 344 pp 40-44

[3] Greenberg P L, Gordeuk V, Issaragrisil S, Siritanaratkul N, Fucharoen S, and Ribeiro R C 2001 Major Hematologic Diseases in the Developing World - New Aspects of Diagnosis and Management of Thalassemia, Malarial Anemia, and Acute Leukemia American Society of Hematology pp 479-98
[4] Rustam Z, Kamalia A, Hidayat R, Subroto F, and Semendawai A S 2019 Comparison of Fuzzy C-Means, Fuzzy Kernel C-Means, and Fuzzy Kernel Robust C-Means to Classify Thalassemia Data. *International journal on Advance Science Engineering Information Technology (IJASEIT)* 9 pp 1205-10

[5] Gu X and Zeng Y 2002 A Review of the Molecular Diagnosis of Thalassemia *Hematology* 7(4) pp 203–9

[6] Nadira T and Rustam Z 2018 Classification of Cancer Data Using Support Vector Machines with Features Selection on Global Artificial Bee Colony *AIP Conf. Proc.* 2023 020205

[7] Aprilliani U and Rustam Z 2018 Osteoarthritis Disease Prediction Based on Random Forest *ICACISIS Conf.* pp 237-40

[8] Huljanah M, Rustam Z, Utama S, and Siswantining T 2019 Feature Selection using Random Forest Classifier for Predicting Prostate Cancer *IOP Conf. Ser.: Mater. Sci. Eng.* 546 052031

[9] Rustam Z and Saragih G S 2018 Predicting Bank Financial Failures using Random Forest *IWBIS Conf.* pp 81-86

[10] Rustam Z, Sudarsono E, Sarwinda D 2019 Random-Forest (RF) and Support Vector Machine (SVM) Implementation for Analysis of Gene Expression Data in Chronic Kidney Disease (CKD) *IOP Conf. Ser.: Mater. Sci. Eng.* 546 052066

[11] Panca V and Rustam Z 2017 Application of Machine Learning on Brain Cancer Multiclass Classification *AIP Conf. Proc.* 1862 030133

[12] Nagalla R, Pothuganti P, and Pawar D S 2017 Analyzing Gap Acceptance Behavior at Unsignalized Intersections Using Support Vector Machines, Decision Tree and Random Forests *Procedia Computer Science* pp 474-481

[13] Rustam Z and Hartini S 2019 Classification of Breast Cancer using Fast Fuzzy Clustering based on Kernel *IOP Conf. Ser.: Mater. Sci. Eng.* 546 052067

[14] Lestari A W and Rustam Z 2017 Normed kernel function-based fuzzy possibilistic C-means (NKFPCM) algorithm for high-dimensional breast cancer database classification with feature selection is based on Laplacian Score *AIP Conf. Proc.* 1862 030143

[15] Rustam Z and Yaurita F 2018 Insolvency Prediction in Insurance Companies Using Support Vector Machines and Fuzzy Kernel C-Means *J. Phys.: Conf. Ser.* 1028 012118

[16] Rachman A A and Rustam Z 2016 Cancer classification using Fuzzy C-Means with feature selection *ICMSA Conf.* pp 31-34

[17] Hu X and Zhao Y 2019 Block Ciphers Classification Based on Random Forest *J. Phys.: Conf. Ser.* 1168 032015

[18] Breiman L 2001 Random Forests Accessed: 20 August 2019 https://www.stat.berkeley.edu/

[19] Nguyen C, Wang Y, and Nguyen H 2013 Random Forest Classifier Combined with Feature Selection for Breast Cancer Diagnosis and Prognostic *JbiSE* 6 pp 551-60

[20] Nithya B and Ilango V 2019 Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction *SN Applied Sciences* 1