Differential Diagnosis of β -Thalassemia Trait from Iron Deficiency Anemia: Application of Bayesian Decision Tree

Mina Jahangiri

Ahvaz Jundishapur University of Medical Sciences: Ahvaz Jondishapour University of Medical Sciences

Fakher Rahim

Ahvaz Jundishapur University of Medical Sciences: Ahvaz Jondishapour University of Medical Sciences

Najmaldin Saki

Ahvaz Jundishapur University of Medical Sciences: Ahvaz Jondishapour University of Medical Sciences

Amal Saki Malehi (✉ amalsaki@gmail.com)

Ahvaz Jondishapour University of Medical Sciences Faculty of Public Health <https://orcid.org/0000-0002-0952-5861>

Research

Keywords: Bayesian Decision Tree, diagnosis, iron deficiency anemia (IDA), β -thalassemia trait (β TT)

Posted Date: January 29th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-157443/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Differential Diagnosis of β -Thalassemia Trait from Iron Deficiency
Anemia: Application of Bayesian Decision Tree**

Mina Jahangiri¹, Fakher Rahim², Najmaldin Saki², Amal Saki Malehi^{1,2}

**1- Department of Biostatistics and Epidemiology, Faculty of Public Health,
Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran
2-Thalassemia & Hemoglobinopathy Research Center, Research
Institute of Health, Ahvaz Jundishapur University of Medical Sciences,
Ahvaz, Iran**

**Correspondence: Amal Saki Malehi, Department of Biostatistics and
Epidemiology, Faculty of Public Health, Ahvaz Jundishapur University of
Medical Sciences, Ahvaz, Iran. Email: amalsaki@gmail.com**

Abstract:

Background: Several discriminating techniques have been proposed to discriminate between β -thalassemia trait (β TT) and iron deficiency anemia (IDA) so far. These discrimination techniques are important clinically, but they are challenging and normally difficult; so if a patient with IDA is diagnosed as β TT, then it is deprived of iron therapy. This study is the first application of the Bayesian tree-based method for differential diagnosis of β TT from IDA.

Method: In this study, 907 patients were enrolled with the ages over 18-year-old with microcytic anemia. Bayesian Logit Treed (BLTREED) has been used to discriminate β TT from IDA.

Results: Mean corpuscular volume (MCV) was found as the main predictor in diagnostic discrimination. BLTREED model showed high sensitivity (96%), specificity (93%), accuracy (95%), Youden's index (89), as well as positive and negative predictive values in the differential diagnosis of β TT from IDA. Also, AUC revealed a more precise classification with an area under the curve value of 0.98.

Conclusions: BLTREED model showed excellent diagnostic accuracy for differentiating β TT from IDA. In addition, understanding tree-based methods are easy and need not a statistical experience, so this advantage can help physicians in making the right clinical decision. Thus, we suggest the using of the BLTREED model as a powerful method in data mining techniques in order to develop sensitive and accurate diagnostic methods for for discriminating between these two anemia disorders.

Keywords: Bayesian Decision Tree, diagnosis, iron deficiency anemia (IDA), β -thalassemia trait (β TT)

Highlight

- 1- To the best of our knowledge this study is the first application of Bayesian tree-based method for differential diagnosis of β TM from IDA.
- 2- We propose an automatic detection model of beta-Thalassemia carriers based on a Bayesian tree-based method.
- 3- The proposed model will support medical decisions for differential diagnosis of β TM from IDA to avoid much more expensive, time-consuming laboratory tests especially in countries with limited resources or poor health services.

1. Introduction:

Iron deficiency anemia (IDA) and β -thalassemia trait (β TT) are the two most common hypochromic microcytic anemia. β TT is more prevalent in the Mediterranean region, in specific geographical areas, including the Caspian Sea and Persian Gulf regions, the 10% prevalence was reported (1). To prevent iron overload and its complications caused by misdiagnosis and inaccurate treatment, and also determining the necessity of prenatal investigations for hemoglobin chain disorders, it is important to differential β TT from IDA (2). Hemoglobin electrophoresis, serum iron and ferritin levels are considered to make a definitive differential diagnosis between β TT and IDA (3-5). However, to reduce costs related to diagnostic workup, various major studies have been conducted to propose appropriate discrimination indices to distinct between β TT and IDA. These indices have been defined to quickly discriminate between IDA and β TT and avoid more time-consuming and expensive methods. Mentzer (6), Shine and Lal (7), England and Fraser (8), RBC (9), Srivastava (10), Ricerca (11), Green and

King (12), Bessman (RDW) (13), Das Gupta (14), Jayabose (RDWI) (15), Telmissani-MCHD (16), Telmissani-MDHL (16), Huber-Herklotz (17), Kerman I (18), Kerman II (18), Sirdah (19), Ehsani (20), Keikhaei (21), Nishad (22), Wongprachum (23), Sehgal (24), Pornprasert (25), Sirachainan (26), Bordbar (27), Matos and Carvalho (28), Janel (11T) (29), CRUISE Index (30), Index26 (30) are hematological discrimination indices used for discriminating between the IDA and β TT. However, these indices obtained empirically and have an inconsistent performance for differential diagnosis of β TT and IDA in the same patient (31). Sometimes the same indices showed different discrimination power in varied age groups (32, 33).

Recently, the accessibility of powerful statistical software has provided the application of data mining techniques for health-related data. Many studies have been proposed advance statistical methods and data mining techniques such as decision Trees methods (68) for differential diagnostic between β TT and IDA to avoid much more expensive, time-consuming, and complicated laboratory procedures and non-satisfactory hematological indices in discriminating between β TT and IDA (34-37, 39, 69, 70). (34-39). Urrechaga, Aguirre, and Izquierdo (38) used multivariable discriminant analysis for differential diagnosis of microcytic anemia.. Wongseree et.al (39) implemented neural network and genetic programming for thalassemia classification. Dogan and Turkoglu (36) proposed a decision tree for detecting iron deficiency anemia from hematology parameters.

Setsirichok (34) evaluated the classification of blood characteristics by a C4.5 decision tree, a naive Bayes classifier and a multilayer perceptron for classifying eighteen classes of thalassemia abnormality. Jahangiri *et al.* (37) used classic decision-tree-based methods for constructing a differential diagnosis scheme and investigating the performance of several tree- based methods for the differential diagnosis of β TT from

1 IDA. Decision Trees have advantages over traditional statistical methods like
2 discriminant analysis, generalized linear models (GLMs) and survival analysis. The
3 main advantage of tree-based methods is tree structure that makes it easy to interpret
4 the clinical data and to be accepted by medical researchers and clinicians. But these
5 methods suffer from greediness problem and this problem have disadvantages like:
6 limit the exploration of tree space, dependence future splits to previous splits, generate
7 optimistic error rates and the inability of the search to find a global optimum (71).
8 Bayesian tree approaches are proposed to solve the greediness problem of tree-
9 based methods. Also, these Bayesian approaches can quantify uncertainty and these
10 approaches explore the tree space more than classic tree approaches. Bayesian
11 approaches combine prior information with observations unlike classic tree methods
12 (these methods use only observations for data analysis). These Bayesian approaches
13 define prior distributions on the components of classic tree methods and then use
14 stochastic search algorithms through Markov chain Monte Carlo (MCMC) algorithms
15 for exploring tree space (72-78). Bayesian tree-based methods have been developed
16 since it can be account more sensibly and comprehensively for uncertainty than
17 frequentist methods.

18 In this paper, a Bayesian tree-based method was proposed for the differential
19 diagnosis of β TT from IDA based on simple laboratory test results.

20 21 2. Material and methods

22 23 2.1 .Criteria For Selecting patient Groups

24 In this study, a total of 907 patients aged over 18 years old diagnosed with IDA or β TT
25 were selected to develop new discriminating indices. Hematological parameters like

Hb (Hemoglobin), Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), Red Blood Cell Distribution Width (RDW), Mean Corpuscular Hemoglobin Concentration (MCHC), and Red Blood Cell count (RBC) were measured by using Sysmex kx-21 automated hematology analyzer.

2.2. Inclusion criteria:

In the IDA group, patients had hemoglobin (Hb) levels less than 12 and 13 g/dL for women and men, respectively. Mean corpuscular hemoglobin (MCH) and Mean corpuscular volume (MCV) were below 80 fL and 27 pg for both sexes, respectively, and for men, ferritin of <28 ng/mL was considered as IDA. In the β TT group, patients had a MCV value below 80 fL. Patients with HbA2 levels of >3.5% were considered as β TT carriers.

2.3. Exclusion criteria:

For the IDA group, patients who had mutations associated with α TT (3.7, 4.2, 20.5, MED, SEA, THAI, FIL, and Hph) were excluded so, individuals presenting the two diseases simultaneously were not selected. For the β TT group, patients with α TT confirmed by presence of mutations in molecular analysis were excluded. All patients with malignancies or inflammatory/infectious diseases diagnosed based on clinical data and personal information obtained from medical records were also excluded.

In addition, pregnant women with severe anemia (Hb < 8 g/dl) and anemia due to chronic disease or other hemoglobinopathies as well as the simultaneous development of IDA and β TT were excluded.

2.4. Ethical consideration

This study was approved and supported by Ethical committee affiliated by the Ahvaz Jundishapur University of Medical Sciences (AJUMS), Ahvaz, Iran.

2.5. Methods

CBC analysis of EDTA-K2 anti-coagulated blood samples was performed using the Sysmex kx-21 automated hematology analyzer (Japan) to measure differential parameters, including Hb, MCV, MCH, and RDW. Also, HbA2 (Hemoglobin A2), Serum Iron, TIBC (Total Iron-Binding Capacity) and Serum Ferritin measured for all patients. The patients with HbA2 > 3.5% diagnosed as β TT, that 537 patients had this condition and 370 patients with serum Iron < 50 μ g/dl, TIBC > 450 μ g/dl and serum Ferritin < 12 μ g/dl diagnosed as IDA. Patients who had conditions of both β TT and IDA groups and or none of the diagnosis conditions of β TT and IDA groups were excluded from the study.

2.6. Statistical Analysis

Decision trees methods are defined as machine-learning methods for constructing prediction models from data and could provide a solution for constructing the diagnostic test (79, 80). These methods are valuable tools in data mining techniques. Tree-based models including nonparametric models and do not need any assumptions about the functional form of the data.

One of the advantages of these methods is the graphical presentation of results that make them easy to interpret and no need for statistical experience for understanding result of models (51, 81-83). Tree-based models also were constructed based on Bayesian algorithms. The Bayesian approach of CART model (BCART) with defining a prior distribution was proposed by Chipman et.al in 1998 (73). Bayesian Logit Treed

(BLTREED) model as an extension of BCART was also developed by Chipman *et al.* by fitting a logistic regression model for data prediction in the terminal nodes (74, 84).

2.6.1. Bayesian Logit Treed (BLTREED) Model

Bayesian approach (BCART) for CART model was constructed by defining a prior distribution on the pair of components of this model namely (θ, T) ; T is a binary tree with $\mathcal{K}$ terminal nodes or tree with size $\mathcal{K}$ and $\theta = (\theta_1, \theta_2, \dots, \theta_{\mathcal{K}})$ is parameters set in the terminal nodes ($\theta_i = p_{ij}$, $i=1, \dots, \mathcal{K}$, $j=1, \dots, N$: the number of distinct classes of the response variable and p_{ij} shows the probability of j th class of response variable in i th terminal node). It can be shown that the joint posterior distribution of parameters and tree structure were as following equation

$$\frac{(\theta|T) p(T)}{p(\theta, T)} \quad (2) = p$$

Where $p(T)$ and $p(\theta|T)$ shows the prior distribution for tree and parameters in terminal nodes respectively. In each Bayesian approach, prior distributions define as unknown, so in this Bayesian approach tree structure and parameters in terminal nodes were considered as unknowns (73). BCART was extended by Chipman *et al.* (2002,2003) by fitting a parametric model such as logistic regression model for data prediction and describe the conditional distribution of $Y|X$ in each terminal nodes (74, 84). In BLTREED model the conditional distribution of $Y|X$ unlike BCART model depends on X ($Y|X \sim f(Y|X, \theta_i)$) and also by fitting sophisticated model at terminal nodes (by fitting logistic regression model for data prediction in each terminal nodes), smaller trees and

more interpretable were generated. In BLTREED model, one subset of X can use to generate the tree and another subset can use for fit models in terminal nodes (these subsets can be joint and or disjoint). In this Bayesian approach $\theta_i = B_i$ shows the set of regression coefficients for the logistic model fitted in i th terminal node.

The recursive stochastic process using a tree-generating stochastic process for tree growing ($p(T)$)

is as follow (73, 74):

1- Start from T that has only a root node (terminal node η).

2- Calculate the probability for splitting node η as follow:

$$P_{\text{split}} = \alpha(1 + d_{\eta})^{-\beta}$$

Where, d_{η} is the depth of the node η , α is the base probability of tree growth of splitting a node, and β is the rate which determine the propensity to split decreases with increased tree size.

Actually (α & β) are parameters that control the shape and size of trees and these parameters provide a penalty to avoid over-fitting model.

3- If the node η splits to left and right nodes according to the distribution of $p_{\text{RULE}}(\rho|\eta, T)$, then let T as the newly created tree from step 3 and re-apply steps 2 and 3 to the new children nodes.

BLTREED model was fitted based on standardized data, so same prior can be used independently for parameters in the terminal nodes and they were considered as a multivariate normal distribution with zero mean and variance matrix proportional to the identity for these parameters (74, 84).

Posterior distribution function $p(T|X, y)$ was computed with combining the marginal likelihood function $p(Y|X, T)$ and tree prior $p(T)$ as follows:

$$p(T|X, y) \propto p(y|X, T) p(T) \quad (3)$$

Where $p(Y|X, T)$ is as follow:

$$p(y|X, T) = \int p(y|X, \theta, T) p(\theta|T) d\theta = \prod_{i=1}^K \int \prod_{h=1}^{n_i} p(y_{ih}|x_{ih}, B_i) p(B_i) dB_i \quad (4)$$

Which $p(y|X, \theta, T)$, (y_{ih}, x_{ih}) and n_i show the data likelihood function, observed values for h_{th} observation in i_{th} node and the number of observations in i_{th} node, respectively.

The integral of equation 4 hasn't closed form, so Laplace approximation was used to solve it (74, 84).

Chipman *et al.* (73, 74) utilize a Metropolis-Hastings algorithm to simulate equation (3) for finding trees with high posterior distribution. The Metropolis-Hastings algorithm simulates a Markov chain sequence of trees namely $T^0, T^1, T^2, \dots$

The simulation algorithm was implemented with multiple restarts for reasons that mentioned in Chipman *et al.* (73, 74).

2.7. Data Analysis

BLTREED models were fitted by using predictor variables such as hemoglobin (Hb), mean cell volume (MCV), mean cell hemoglobin (MCH) and red cells distribution width (RDW) for differential diagnosis of β TT from IDA.

BLTREED model fitted using 8 restarts with 6000 iterations per restart and use a prior standard deviation of 20 for the logit coefficients (84). For determining the pair of (α, β) , BLTREED model was fitted with two choices 0.5 and 0.95 for α parameter, and four choices for β (a range 0.5- 2 by step 0.5), then selected the pair of (α, β) that generate the best tree with smallest FNR.

Differential performance of the Bayesian classification tree was evaluated using criteria such as sensitivity (TPR), specificity (TNR), false negative rate (FNR) and false positive rate (FPR), positive predictive value (PPV) and negative predictive value (NPV), positive likelihood ratio (PLR) and negative likelihood ratio (NLR), accuracy, Youden's index, area under curve (AUC) and F-measure. AUC can be interpreted as an overall performance measure in the classification of the tree models. Criteria such as F-measure, Youden's index, accuracy, PLR, NLR (a good diagnostic test or has $NLR < 0.1$ and a good diagnostic test has $PLR > 10$) and AUC take both sensitivity and specificity into consideration, so they can present the performance of model more accurately than other criteria. $P < 0.05$ was considered to be statistically significant.

2.8. Software

Data were analyzed by free software (<http://gsbwww.uchicago.edu/fac/robert.mcculloch/research/code/CART/index.html>) based on Chipman et al. (2002) that was developed for fitting BTREED model, R 3.0.3 used for compute performance measures (ePiR and pROC), split data to training data set and test data set (caTools package).

3. Results

A total of 537 patients were diagnosed as β TT including 299 (56%) women and 238 (44%) men, while 370 patients were diagnosed as IDA included 293 (79%) women and 77 (21%) men. Table 1 shows Mean, standard deviation (SD) of laboratory parameter as predictor variables across the type of hypochromic microcytic anemia (β TT and IDA).

The tree structure of BLTREED model was shown in figures 1. The first split of the tree was based on MCV, it showed that MCV is an important predictor in differentiation between the types of hypochromic microcytic anemia. Another predictor which used as second splitting variable in tree structure was HB. According to the presented tree, four homogenous sub-groups were extracted from data which obtained four diagnostic discrimination rules for differentiating between β TT and IDA (Table 2). This classifying scheme showed that values of $MCV \leq 72.70$ screening the β TT patients.

Predictive performance of the model in differentiation between β TT and IDA calculated based on confusion matrix (Table 3). The obtained tree showed the highly TPR, TNR, PPV, NPV, Youden's Index, accuracy and F-measure in differentiation between β TT and IDA (Table 4).

In addition, the model has $NLR < 0.1$ and it could be concluded that BLTREED has good diagnostic accuracy for discriminating the patients. Table 5 shows AUCs of the model from ROC analysis that were statistically significant ($p < 0.001$) and showed the excellent diagnose accuracy in differentiation between the types of hypochromic microcytic anemia.

4. Discussion:

In this paper, we used BLTREED model as a Bayesian decision tree as the differential diagnostic tool for thalassemia diagnosis. This is a first study that uses BLTREED

1 model in the hematological data. The Bayesian decision tree used to solve uncertainly
2 problems of conventional tree-based methods (74, 84, 85). This model was
3 implemented by using Hb, MCV, MCH and RDW as independent variables. Based on
4 our result, MCV and Hb were main predictor parameters in differential diagnostic and
5 it showed that the patient with β TT has lower values of MCV.

6 In previous studies that used the different conventional decision trees for differential
7 diagnosis β TT from IDA, the first split of all algorithms was based on MCV and they
8 also concluded that MCV was an important predictor variable in discrimination of IDA
9 and β TT (35, 37). The performance of BLTREED model that evaluated by using
10 sensitivity, specificity, false negative and positive rate, positive and negative predictive
11 value, exhibited the high performance of the differential diagnosis of β TT from IDA. In
12 addition, positive likelihood ratio and negative likelihood ratio, accuracy,
13 Youden's index and F-measure showed that BLTREED has good diagnostic accuracy
14 for discriminating the patients. It was truly classified 96% of β TT patient. Furthermore,
15 AUC as an index of overall performance showed excellent and significant accuracy
16 (99, 98) in training and test data, respectively in differential diagnostic of β TT and IDA.
17 Other studies that used different data mining techniques and decision trees based on
18 frequentist approach of fitting revealed the high performance and accuracy but lower
19 than our result (34-37, 70). BLTREED model improves the classification performance
20 by solving the uncertainty of previous models (74, 84).

21 The diagnostic performance of BLTREED was better than other discrimination
22 methods (classification trees or hematological discrimination indices) in past studies
23 for differentiating β TT from IDA. These studies are as follows: Setsirichok et al. (2012)
24 used C4.5 decision tree, naïve Bayes (NB) classifier and multilayer perceptron (MLP)
25 for classifying eighteen classes of thalassemia abnormality (34). Bellinger et al. (2015)

used classification algorithms like J48 decision tree, support vector machines (SVM), k-nearest neighbors (k-NN), MLP and NB for differentiating between β TM, IDA and co-occurrence of these disorders (70). AlAgha *et al.* (2018) compared the diagnostic performance of different classification algorithms such as J48, k-NN, artificial neural networks (ANN) and NB for classifying β -thalassemia carriers (86). Jahangiri *et al.* (2017) utilized classification tree algorithms such as CHAID, E-CHAID, CART, QUEST, GUIDE and CRUISE for differential diagnosis of β TT from IDA. They indicated that CRUISE algorithm has the best diagnostic performance with similar to and the present study, but this classic algorithm uses greedy algorithm for tree generating and cannot explore the tree space more than Bayesian tree approaches. Also, many studies compared the diagnostic performance of hematological discrimination indices and BLTREED showed better performance in comparison to them (19-22, 26, 28-33, 87-102).

5. Conclusion

In the present study, BLTREED model showed excellent diagnostic accuracy for differentiating β TT from IDA. According to the advantages of Bayesian tree-based methods like generating a small and more interpretable tree, and lack of uncertainty of different conventional decision trees, this method can be helpful along with other laboratory parameters for discriminating between these two anemia disorders. Also, understanding tree-based methods are easy and need not a statistical experience, so this advantage can help physicians in making the right clinical decision.

Abbreviations

1 β TT : β -thalassemia trait; IDA: Iron deficiency anemia; MCV: Mean Corpuscular
2 Volume ; MCH: Mean Corpuscular Hemoglobin ; RDW: Red Blood Cell Distribution
3 Width ; MCHC: Mean Corpuscular Hemoglobin Concentration; RBC: Red Blood Cell;
4 BLTREED: Bayesian Logit Treed; TPR: sensitivity; TNR: specificity; FNR: false
5 negative rate; FPR: false positive rate; NPV: negative predictive value, PPV: positive
6 predictive value; PLR: positive likelihood ratio; NLR:negative likelihood ratio.

7 8 Acknowledgements

9 This paper is part of thesis of Mina Jahangiri, MSc student of Biostatistics.

10 Funding

11 This work was financially supported by grant no. U-95095 from vice chancellor for
12 Research Affairs of Ahvaz Jundishapur University of Medical Sciences.

13 14 Availability of data and materials

15 The datasets used and/or analyzed during the current study available from the
16 corresponding author on reasonable request.

17 18 Authors' contributions

19 ASM and MJ: Conception and design; Analysis and interpretation of the data; Drafting
20 of the article. FR, NS: Conception and design; Collection and assembly of data,
21 Drafting of the article. All authors approved the final version of the article for
22 submission.

23 24 Ethics approval and consent to participate

This study was approved by the Ethics Committee of Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran. (IR.AJUMS.REC.1395.456)

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References:

1. Batebi A, Pourreza A, Esmailian R. Discrimination of beta-thalassemia minor and iron deficiency anemia by screening test for red blood cell indices. Turkish Journal of Medical Sciences. 2012;42(2):275-80.

- 1 2. Hallberg L. Iron requirements. Biological trace element research.
2 1992;35(1):25-45.
- 3 3. Goddard AF, James MW, McIntyre AS, Scott BB. Guidelines for the
4 management of iron deficiency anaemia. Gut. 2011;gut. 2010.228874.
- 5 4. Mosca A, Paleari R, Ivaldi G, Galanello R, Giordano P. The role of haemoglobin
6 A2 testing in the diagnosis of thalassaemias and related haemoglobinopathies.
7 Journal of Clinical Pathology. 2009;62(1):13-7.
- 8 5. Oliveri N. The beta-thalassemyias. N Engl J Med. 1999;341(2):99-109.
- 9 6. Mentzer W. Differentiation of iron deficiency from thalassaemia trait. The
10 Lancet. 1973;301(7808):882.
- 11 7. Shine I, Lal S. A strategy to detect β -thalassaemia minor. The Lancet.
12 1977;309(8013):692-4.
- 13 8. England J, Fraser P. Differentiation of iron deficiency from thalassaemia trait by
14 routine blood-count. The Lancet. 1973;301(7801):449-52.
- 15 9. Klee GG, Fairbanks VF, Pierre RV, O'sullivan MB. Routine erythrocyte
16 measurements in diagnosis of iron-deficiency anemia and thalassemia minor.
17 American Journal of Clinical Pathology. 1976;66(5):870-7.
- 18 10. Srivastava P, Bevington J. Iron deficiency and/or Thalassaemia trait. The
19 Lancet. 1973;301(7807):832.
- 20 11. Ricerca B, Storti S, d'Onofrio G, Mancini S, Vittori M, Campisi S, et al.
21 Differentiation of iron deficiency from thalassaemia trait: a new approach.
22 Haematologica. 1986;72(5):409-13.
- 23 12. Green R, King R. A new red cell discriminant incorporating volume dispersion
24 for differentiating iron deficiency anemia from thalassemia minor. Blood cells.
25 1989;15(3):481-95.

- 1 13. Bessman JD, Feinstein D. Quantitative anisocytosis as a discriminant between
2 iron deficiency and thalassemia minor. *Blood*. 1979;53(2):288-93.
- 3 14. Gupta AD, Hegde C, Mistri R. Red cell distribution width as a measure of
4 severity of iron deficiency in iron deficiency anemia. *Indian J Med Res*. 1994;100:177-
5 83.
- 6 15. Jayabose S, Giamelli J, LevondogluTugal O, Sandoval C, Ozkaynak F,
7 Visintainer P. # 262 Differentiating iron deficiency anemia from thalassemia minor by
8 using an RDW-based index. *Journal of Pediatric Hematology/Oncology*.
9 1999;21(4):314.
- 10 16. TELMISSANI OA, KHALIL S, ROBERTS GT. Mean density of hemoglobin per
11 liter of blood: a new hematologic parameter with an inherent discriminant function.
12 *Laboratory Hematology*. 1999;5:149-52.
- 13 17. Huber AR, Ottiger C, Risch L, Regenass S, Hergersberg M, Herklotz R, editors.
14 *Thalassemie-syndrom: klinik und diagnose*. Schweiz Med Forum; 2004.
- 15 18. KOHAN N, Ramzi M. Evaluation of sensitivity and specificity of Kerman index I
16 and II in screening beta thalassemia minor. 2008.
- 17 19. Sirdah M, Tarazi I, Al Najjar E, Al Haddad R. Evaluation of the diagnostic
18 reliability of different RBC indices and formulas in the differentiation of the β -
19 thalassaemia minor from iron deficiency in Palestinian population. *International*
20 *Journal of Laboratory Hematology*. 2008;30(4):324-30.
- 21 20. Ehsani M, Shahgholi E, Rahiminejad M, Seighali F, Rashidi A. A new index for
22 discrimination between iron deficiency anemia and beta-thalassemia minor: results in
23 284 patients. *Pakistan journal of biological sciences: PJBS*. 2009;12(5):473-5.
- 24 21. Keikhaei B. A new valid formula in differentiating iron deficiency anemia from
25 β -thalassemia trait. *Pakist J Med Sci*. 2010;26:368-73.

- 1 22. Nishad AAN, Pathmeswaran A, Wickremasinghe A, Premawardhena A. The
2 Thal-index with the BTT prediction. exe to discriminate β -thalassaemia traits from
3 other microcytic anaemias. 2012.
- 4 23. Wongprachum K, Sanchaisuriya K, Sanchaisuriya P, Siridamrongvattana S,
5 Manpeun S, Schlep FP. Proxy indicators for identifying iron deficiency among anemic
6 vegetarians in an area prevalent for thalassemia and hemoglobinopathies. *Acta*
7 *haematologica*. 2012;127(4):250-5.
- 8 24. Dharmani P, Sehgal K, Dadu T, Mankeshwar R, Shaikh A, Khodaiji S.
9 Developing a new index and its comparison with other CBC-based indices for
10 screening of beta thalassemia trait in a tertiary care hospital. *International Journal of*
11 *Laboratory Hematology*. 2013;35:118.
- 12 25. Pornprasert S, Panya A, Punyamung M, Yanola J, Kongpan C. Red cell indices
13 and formulas used in differentiation of β -thalassemia trait from iron deficiency in Thai
14 school children. *Hemoglobin*. 2014;38(4):258-61.
- 15 26. Sirachainan N, Iamsirirak P, Charoenkwan P, Kadegasem P,
16 Wongwerawattanakoon P, Sasanakul W, et al. New mathematical formula for
17 differentiating thalassemia trait and iron deficiency anemia in thalassemia prevalent
18 area: a study in healthy school-age children. *Southeast Asian Journal of Tropical*
19 *Medicine and Public Health*. 2014;45(1):174.
- 20 27. Bordbar E, Taghipour M, Zucconi BE. Reliability of different RBC indices and
21 formulas in discriminating between β -thalassemia minor and other microcytic
22 hypochromic cases. *Mediterranean journal of hematology and infectious diseases*.
23 2015;7(1).

28. Matos JF, Dusse L, Borges KB, de Castro RL, Coura-Vital W, Carvalho MdG. A new index to discriminate between iron deficiency anemia and thalassemia trait. *Revista brasileira de hematologia e hemoterapia*. 2016;38(3):214-9.
29. Janel A, Roszyk L, Rapatel C, Mareynat G, Berger MG, Serre-Sapin AF. Proposal of a score combining red blood cell indices for early differentiation of beta-thalassemia minor from iron deficiency anemia. *Hematology*. 2011;16(2):123-7.
30. Jahangiri M, Rahim F, Malehi AS. Diagnostic performance of hematological discrimination indices to discriminate between β thalassemia trait and iron deficiency anemia and using cluster analysis: Introducing two new indices tested in Iranian population. *Scientific Reports*. 2019;9(1):1-13.
31. Vehapoglu A, Ozgurhan G, Demir AD, Uzuner S, Nursoy MA, Turkmen S, et al. Hematological indices for differential diagnosis of beta thalassemia trait and iron deficiency anemia. *Anemia*. 2014;2014.
32. Rahim F, Keikhaei B. Better differential diagnosis of iron deficiency anemia from beta-thalassemia trait. *Turk J Hematol*. 2009;26(3):138-45.
33. Hoffmann JJ, Urrechaga E, Aguirre U. Discriminant indices for distinguishing thalassemia and iron deficiency in patients with microcytic anemia: a meta-analysis. *Clinical Chemistry and Laboratory Medicine (CCLM)*. 2015;53(12):1883-94.
34. Setsirichok D, Piroonratana T, Wongseeree W, Usavanarong T, Paulkhaolarn N, Kanjanakorn C, et al. Classification of complete blood count and haemoglobin typing data by a C4. 5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening. *Biomedical Signal Processing and Control*. 2012;7(2):202-12.
35. Elshami EH, Alhalees AM, editors. *Automated Diagnosis of Thalassemia Based on DataMining Classifiers*. The International Conference on Informatics and

Applications (ICIA2012); 2012: The Society of Digital Information and Wireless Communication.

36. Dogan S, Turkoglu I. Iron-deficiency anemia detection from hematology parameters by using decision trees. International Journal of Science & Technology. 2008;3(1):85-92.

37. Jahangiri M, Khodadi E, Rahim F, Saki N, Saki Malehi A. Decision-tree-based methods for differential diagnosis of β -thalassemia trait from iron deficiency anemia. Expert Systems. 2017.

38. Urrechaga E, Aguirre U, Izquierdo S. Multivariable discriminant analysis for the differential diagnosis of microcytic anemia. Anemia. 2013;2013.

39. Wongseree W, Chaiyaratana N, Vichittumaros K, Winichagoon P, Fucharoen S. Thalassaemia classification by neural networks and genetic programming. Information Sciences. 2007;177(3):771-86.

40. Ilgen MA, Downing K, Zivin K, Hoggatt KJ, Kim HM, Ganoczy D, et al. Identifying subgroups of patients with depression who are at high risk for suicide. The Journal of clinical psychiatry. 2009;70(11):1495.

41. Sledjeski EM, Dierker LC, Brigham R, Breslin E. The use of risk assessment to predict recurrent maltreatment: A classification and regression tree analysis (CART). Prevention science. 2008;9(1):28-37.

42. Buntinx F, Truyen J, Embrechts P, Moreel G, Peeters R. Evaluating patients with chest pain using classification and regression trees. Family practice. 1992;9(2):149-53.

43. Tsien CL, Fraser H, Long WJ, Kennedy RL. Using classification tree and logistic regression methods to diagnose myocardial infarction. Studies in health technology and informatics. 1998(1):493-7.

- 1 44. Goldman L, Weinberg M, Weisberg M, Olshen R, Cook EF, Sargent RK, et al.
2 A computer-derived protocol to aid in the diagnosis of emergency room patients with
3 acute chest pain. *New England journal of medicine*. 1982;307(10):588-96.
- 4 45. Selker HP, Griffith JL, Patil S, Long WJ, d'Agostino R. A comparison of
5 performance of mathematical predictive methods for medical diagnosis: identifying
6 acute cardiac ischemia among emergency department patients. *Journal of*
7 *investigative medicine: the official publication of the American Federation for Clinical*
8 *Research*. 1995;43(5):468-76.
- 9 46. Stewart PW, Stamm JW. Classification tree prediction models for dental caries
10 from clinical, microbiological, and interview data. *Journal of dental research*.
11 1991;70(9):1239-51.
- 12 47. Lieu TA, Quesenberry CP, Sorel ME, Mendoza GR, Leong AB. Computer-
13 based models to identify high-risk children with asthma. *American Journal of*
14 *Respiratory and Critical Care Medicine*. 1998;157(4):1173-80.
- 15 48. Malehi AS. Diagnostic Classification Scheme in Iranian Breast Cancer Patients
16 using a Decision Tree. *Asian Pacific journal of cancer prevention: APJCP*.
17 2013;15(14):5593-6.
- 18 49. Moisen G. *Classification and regression trees*. 2008.
- 19 50. Thuiller W, Araújo MB, Lavorel S. Generalized models vs. classification tree
20 analysis: predicting spatial distributions of plant species at different scales. *Journal of*
21 *Vegetation Science*. 2003;14(5):669-80.
- 22 51. De'ath G, Fabricius KE. *Classification and regression trees: a powerful yet*
23 *simple technique for ecological data analysis*. *Ecology*. 2000;81(11):3178-92.

- 1 52. Edwards TC, Cutler DR, Zimmermann NE, Geiser L, Moisen GG. Effects of
2 sample survey design on the accuracy of classification tree models in species
3 distribution models. *Ecological Modelling*. 2006;199(2):132-41.
- 4 53. Debeljak M, Džeroski S, Jerina K, Kobler A, Adamič M. Habitat suitability
5 modelling for red deer (*Cervus elaphus* L.) in South-central Slovenia with classification
6 trees. *Ecological Modelling*. 2001;138(1):321-30.
- 7 54. Miller J, Franklin J. Modeling the distribution of four vegetation alliances using
8 generalized linear models and classification trees with spatial dependence. *Ecological*
9 *Modelling*. 2002;157(2):227-47.
- 10 55. Che D, Hockenbury C, Marmelstein R, Rasheed K. Classification of genomic
11 islands using decision trees and their ensemble algorithms. *BMC genomics*.
12 2010;11(2):1.
- 13 56. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale,
14 application, and characteristics of classification and regression trees, bagging, and
15 random forests. *Psychological Methods*. 2009;14(4):323.
- 16 57. Chen X, Wang M, Zhang H. The use of classification trees for bioinformatics.
17 *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2011;1(1):55-
18 63.
- 19 58. Marshall RJ. The use of classification and regression trees in clinical
20 epidemiology. *Journal of clinical epidemiology*. 2001;54(6):603-9.
- 21 59. Bachur RG, Harper MB. Predictive model for serious bacterial infections among
22 infants younger than 3 months of age. *Pediatrics*. 2001;108(2):311-6.
- 23 60. Carmelli D, Zhang H, Swan GE. Obesity and 33-year follow-up for coronary
24 heart disease and cancer mortality. *Epidemiology*. 1997:378-83.

- 1 61. El-Solh AA, Sikka P, Ramadan F. Outcome of older patients with severe
2 pneumonia predicted by recursive partitioning. *Journal of the American Geriatrics*
3 *Society*. 2001;49(12):1614-21.
- 4 62. Kuchibhatla M, Fillenbaum GG. Assessing risk factors for mortality in elderly
5 White and African American people: implications of alternative analyses. *The*
6 *Gerontologist*. 2002;42(6):826-34.
- 7 63. Nelson LM, Bloch DA, Longstreth W, Shi H. Recursive partitioning for the
8 identification of disease risk subgroups: a case-control study of subarachnoid
9 hemorrhage. *Journal of clinical epidemiology*. 1998;51(3):199-209.
- 10 64. Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on
11 differential effect search—a recursive partitioning method for establishing response to
12 treatment in patient subpopulations. *Statistics in Medicine*. 2011;30(21):2601-21.
- 13 65. Loh WY, He X, Man M. A regression tree approach to identifying subgroups
14 with differential treatment effects. *Statistics in Medicine*. 2015;34(11):1818-33.
- 15 66. Su X, Tsai C-L, Wang H, Nickerson DM, Li B. Subgroup analysis via recursive
16 partitioning. *Journal of Machine Learning Research*. 2009;10(Feb):141-58.
- 17 67. Li C, Glüer C-C, Eastell R, Felsenberg D, Reid DM, Roux C, et al. Tree-
18 structured subgroup analysis of receiver operating characteristic curves for diagnostic
19 tests. *Academic radiology*. 2012;19(12):1529-36.
- 20 68. Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. *Machine*
21 *Learning*. 1997;29(2-3):131-63.
- 22 69. Maity M, Mungle T, Dhane D, Maiti AK, Chakraborty C. An Ensemble Rule
23 Learning Approach for Automated Morphological Classification of Erythrocytes.
24 *Journal of medical systems*. 2017;41(4):56.

- 1 70. Bellinger C, Amid A, Japkowicz N, Victor H, editors. Multi-label Classification of
2 Anemia Patients. Machine Learning and Applications (ICMLA), 2015 IEEE 14th
3 International Conference on; 2015: IEEE.
- 4 71. Malehi AS, Jahangiri M. Classic and Bayesian Tree-Based Methods.
5 Enhanced Expert Systems: IntechOpen; 2019.
- 6 72. Denison DG, Mallick BK, Smith AF. A bayesian CART algorithm. Biometrika.
7 1998;85(2):363-77.
- 8 73. Chipman HA, George EI, McCulloch RE. Bayesian CART model search.
9 Journal of the American Statistical Association. 1998;93(443):935-48.
- 10 74. Chipman H, George E, McCulloch R. Bayesian treed generalized linear models.
11 Bayesian statistics. 2003;7:323-49.
- 12 75. Chipman HA, George EI, McCulloch RE. Bayesian treed models. Machine
13 Learning. 2002;48(1-3):299-320.
- 14 76. Wu Y, Tjelmeland H, West M. Bayesian CART: Prior specification and posterior
15 simulation. Journal of Computational and Graphical Statistics. 2007;16(1):44-66.
- 16 77. O'Leary RA, Murray JV, Low Choy SJ, Mengersen KL. Expert elicitation for
17 Bayesian classification trees. Journal of Applied Probability & Statistics. 2008;3(1):95-
18 106.
- 19 78. Hu W, O'Leary RA, Mengersen K, Choy SL. Bayesian classification and
20 regression trees for predicting incidence of cryptosporidiosis. PloS one.
21 2011;6(8):e23903.
- 22 79. Breiman L, Friedman J, Olshen RA, Stone CJ. Classification and Regression
23 Trees. New York: Chapman and Hall/CRC; 1984.

- 1 80. Zhang H, H.Singer B. Recursive Partitioning and Applications. Second ed.
2 Bickel P, Diggle P, Fienberg S, Gather U, Olkin I, Zeger S, editors. New York: Springer;
3 2010. 8 p.
- 4 81. Loh WY. Classification and regression trees. Wiley Interdisciplinary Reviews:
5 Data Mining and Knowledge Discovery. 2011;1(1):14-23.
- 6 82. Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and
7 regression tree analysis in public health: methodological review and comparison with
8 logistic regression. Annals of behavioral medicine. 2003;26(3):172-81.
- 9 83. Speybroeck N, Berkvens D, Mfoukou-Ntsakala A, Aerts M, Hens N, Van
10 Huylenbroeck G, et al. Classification trees versus multinomial models in the analysis
11 of urban farming systems in Central Africa. Agricultural Systems. 2004;80(2):133-49.
- 12 84. Moe WW, Chipman H, George EI, McCulloch RE. A Bayesian treed model of
13 online purchasing behavior using in-store navigational clickstream. revising for 2nd
14 review at Journal of Marketing Research. 2002.
- 15 85. Gray JB, Fan G. Classification tree analysis using TARGET. Computational
16 Statistics & Data Analysis. 2008;52(3):1362-72.
- 17 86. AlAgha AS, Faris H, Hammo BH, Ala'M A-Z. Identifying β -thalassemia carriers
18 using a data mining approach: The case of the Gaza Strip, Palestine. Artificial
19 intelligence in medicine. 2018;88:70-83.
- 20 87. Hoffmann JJ, Urrechaga E. Role of RDW in mathematical formulas aiding the
21 differential diagnosis of microcytic anemia. Scandinavian Journal of Clinical and
22 Laboratory Investigation. 2020:1-6.
- 23 88. Miri-Moghaddam E, Sargolzaie N. Cut off determination of discrimination
24 indices in differential diagnosis between iron deficiency anemia and β -thalassemia

minor. International journal of hematology-oncology and stem cell research.
2014;8(2):27.

89. Nesa A, Tayab MA, Sultana T, Khondker L, Rahman MQ, Karim MA, et al.
RDWI is Better Discriminant than RDW in Differentiation of Iron Deficiency Anaemia
and Beta Thalassaemia Trait. Bangladesh Journal of Child Health. 2009;33(3):100-3.

90. Beyan C, Kaptan K, Ifran A. Predictive value of discrimination indices in
differential diagnosis of iron deficiency anemia and beta-thalassemia trait. European
journal of haematology. 2007;78(6):524-6.

91. Ghafouri M, MOSTAAN SL, SHARIFI S, HOSSEINI GL, ATAR CZ. Comparison
of cell counter indices in differentiation of beta thalassemia minor from iron deficiency
anemia. 2006.

92. Demir A, Yarali N, Fisgin T, Duru F, Kara A. Most reliable indices in
differentiation between thalassemia trait and iron deficiency anemia. Pediatrics
International. 2002;44(6):612-6.

93. Schoorl M, Schoorl M, Linssen J, Villanueva MM, NoGuera JAV, Martinez PH,
et al. Efficacy of advanced discriminating algorithms for screening on iron-deficiency
anemia and β -thalassemia trait: a multicenter evaluation. American Journal of Clinical
Pathology. 2012;138(2):300-4.

94. Tripathi N, Soni JP, Sharma PK, Verma M. Role of Haemogram Parameters
and RBC Indices in Screening and Diagnosis of Beta-Thalassemia Trait in Microcytic,
Hypochromic Indian Children. International Journal of Hematological Disorders.
2015;2(2):43-6.

95. Roth IL, Lachover B, Koren G, Levin C, Zalman L, Koren A. Detection of β -
thalassemia carriers by red cell parameters obtained from automatic counters using

1 mathematical formulas. Mediterranean journal of hematology and infectious diseases.
2 2018;10(1).

3 96. Matos JF, Dusse LMSA, Stubbert RVB, Ferreira MR, Coura-Vital W, Fernandes
4 APSM, et al. Comparison of discriminative indices for iron deficiency anemia and β
5 thalassemia trait in a Brazilian population. Hematology. 2013;18(3):169-74.

6 97. Getta HA, Yasseen HA, Said HM. Hi & Ha, are new indices in differentiation
7 between Iron deficiency anemia and beta-Thalassaemia trait. A Study in Sulaimani
8 City-Kurdistan/Iraq IOSR-JDMS. 2015;14(7):67-72.

9 98. Jameel T, Baig M, Ahmed I, Hussain MB, bin Doghaim Alkhamaly M.
10 Differentiation of beta thalassemia trait from iron deficiency anemia by hematological
11 indices. Pakistan journal of medical sciences. 2017;33(3):665.

12 99. Tong L, Kauer J, Wachsmann-Hogiu S, Chu K, Dou H, Smith ZJ. A new red cell
13 index and portable rbc analyzer for screening of iron deficiency and thalassemia minor
14 in a chinese population. Scientific Reports. 2017;7(1):1-10.

15 100. Shen C, Jiang Y-m, Shi H, Liu J-h, Zhou W-j, Dai Q-k, et al. Evaluation of indices
16 in differentiation between iron deficiency anemia and β -thalassemia trait for Chinese
17 children. Journal of Pediatric Hematology/Oncology. 2010;32(6):e218-e22.

18 101. Urrechaga E, Hoffmann JJ. Critical appraisal of discriminant formulas for
19 distinguishing thalassemia from iron deficiency in patients with microcytic anemia.
20 Clinical Chemistry and Laboratory Medicine (CCLM). 2017.

21 102. Jahangiri M, Rahim F, Saki Malehi A, Pezeshki SMS, Ebrahimi M. Differential
22 Diagnosis of Microcytic Anemia, Thalassemia or Iron Deficiency Anemia: A Diagnostic
23 Test Accuracy Meta-Analysis. Modern Medical Laboratory Journal. 2019;3(1):1-14.

Table 1 . Hematological data of study groups presented as mean $\pm$ SD

	β TT (n = 537)	IDA (n=370)	P-value
MCV	62.17 ± 4.14	71.87 ± 6.93	< 0.001
MCH	19.75 ± 1.45	21.85 ± 2.99	< 0.001
Hb	11.20 ± 1.41	10.82 ± 2.43	< 0.001
RDW	15.88 ± 1.43	16.04 ± 2.31	0.94

Table 2. Subgroups extracted for diagnose of β TT and IDA patients by BLTREED model

Subgroup	Conditions	Diagnose label ⁸
A	$MCV \leq 70.70 + HB \leq 12.70$	β TT ⁹
B	$70.70 < MCV \leq 72.70 + HB \leq 12.70$	IDA ¹⁰
C	$HB \leq 12.70 + MCV < 72.70$	β TT
D	$MCV > 72.70$	IDA ¹²

1

2 Table 3. Confusion table of BLTREED model for training data set and test data set

	Training data set			Test data set		
	Predicted			Predicted		
	β TT	IDA	Total	β TT	IDA	Total
Actual	363	13	376	155	6	161
β TT	25	234	259	8	103	111
IDA	388	247	635	163	109	272
Total						

12 Table 4. Sensitivity, specificity, false positive and negative rate, positive and negative
 13 predictive values, accuracy, Youden's index, positive and negative likelihood ratio
 14 and F-measure of BLTREED model in prediction of IDA and β TT groups and their
 15 95% exact confidence interval for training and test data set

Diagnostic Variables	Training data set	Test data set
TPR	97 (94 – 98)	96 (92 – 99)
TNR	90 (86 – 94)	93 (86 – 97)
FNR	3 (2 – 6)	4 (1 – 8)
FPR	10 (6 – 14)	7 (3 – 14)
PPV	94 (91 – 96)	95 (91 – 98)
NPV	95 (91 – 97)	94 (88 – 98)
Youden's Index	87 (80 – 92)	89 (78 – 95)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

Accuracy	94 (92 – 96)	95 (91–97)
LR+	10 (7 –14)	13.36 (7–26)
LR–	0.04 (0.02 – 0.07)	0.04 (0.02 – 0.09)
F–measure	0.96	0.96

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

Table 5. Area under Roc Curve of BTREED model in prediction of IDA and β TT groups for training and test data set (SE: Standard Error of AUC, CI: Confidence Interval)

	Training data set	Test data set
AUC	0.99	0.98
SE	0.003	0.009
95% CI	0.98–0.99	0.96–0.99
P-value	< 0.001	< 0.001

1

2

3

4

Figures

Figure 1

Decision tree for BTREED model ($\alpha= 0.95, \beta= 1$, Log integrated likelihood= -123.423367)