Hindawi
Computational and Mathematical Methods in Medicine
Volume 2021, Article ID 6401105, 10 pages
https://doi.org/10.1155/2021/6401105

Research Article

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

Mina Jahangiri,[1] **Fakher Rahim**,[2] **Najmaldin Saki**,[2] **and Amal Saki Malehi**[2,3]

[1]*Ph.D. Student, Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran*
[2]*Thalassemia & Hemoglobinopathy Research Center, Research Institute of Health, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran*
[3]*Department of Biostatistics and Epidemiology, Faculty of Public Health, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran*

Correspondence should be addressed to Amal Saki Malehi; amalsaki@gmail.com

Received 2 June 2021; Revised 21 September 2021; Accepted 11 October 2021; Published 9 November 2021

Academic Editor: Giovanni D Addio

Objective. Several discriminating techniques have been proposed to discriminate between β-thalassemia trait (βTT) and iron deficiency anemia (IDA). These discrimination techniques are essential clinically, but they are challenging and typically difficult. This study is the first application of the Bayesian tree-based method for differential diagnosis of βTT from IDA. *Method*. This cross-sectional study included 907 patients with ages over 18 years old and a mean (±SD) age of 25 ± 16.1 with either βTT or IDA. Hematological parameters were measured using a Sysmex KX-21 automated hematology analyzer. Bayesian Logit Treed (BLTREED) and Classification and Regression Trees (CART) were implemented to discriminate βTT from IDA based on the hematological parameters. *Results*. This study proposes an automatic detection model of beta-thalassemia carriers based on a Bayesian tree-based method. The BLTREED model and CART showed that mean corpuscular volume (MCV) was the main predictor in diagnostic discrimination. According to the test dataset, CART indicated higher sensitivity and negative predictive value than BLTREED for differential diagnosis of βTT from IDA. However, the CART algorithm had a high false-positive rate. Overall, the BLTREED model showed better performance concerning the area under the curve (AUC). *Conclusions*. The BLTREED model showed excellent diagnostic accuracy for differentiating βTT from IDA. In addition, understanding tree-based methods are easy and do not need statistical experience. Thus, it can help physicians in making the right clinical decision. So, the proposed model could support medical decisions in the differential diagnosis of βTT from IDA to avoid much more expensive, time-consuming laboratory tests, especially in countries with limited recourses or poor health services.

1. Introduction

Iron deficiency anemia (IDA) and β-thalassemia trait (βTT) are the two most common hypochromic microcytic anemia. βTT 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]. The differential between βTT from IDA is crucial for preventing iron overload and related complications caused by misdiagnosis and inaccurate treatment [2].

Differentiation of β-thalassemia trait from iron deficiency anemia is also essential for premarital counseling in developed countries; for patients with microcytic anemia, complete blood count (CBC), in conjunction with hemoglobin variant analysis by high-performance liquid chromatography (HPLC), is interpreted to differentiate iron deficiency

from thalassemia traits. Then, iron studies and molecular testing are also performed. Hemoglobin electrophoresis, serum iron, and ferritin levels are considered to make a definitive differential diagnosis between βTT and IDA [3–5].

However, in low-resource settings where HPLC and molecular testing are not available, different studies proposed discrimination indices to distinct between βTT and IDA. These indices have been defined to quickly discriminate between IDA and βTT and avoid more time-consuming and expensive methods. Mentzer [3], Shine and Lal [4], England and Fraser [5], RBC [6], Srivastava and Bevington [7], Ricerca et al. [8], Green and King [9], Bessman and Feinstein (RDW) [10], Gupta et al. [11], Jayabose et al. (RDWI) [12], Telmissani-MCHD [13], Telmissani-MDHL [13], Huber-Herklotz [14], Kerman I [15], Kerman II [15], Sirdah et al. [16], Ehsani et al. [17], Keikhaei [18], Nishad et al. [19], Wongprachum et al. [20], Dharmani et al. [21], Pornprasert et al. [22], Sirachainan et al. [23], Bordbar et al. [24], Matos et al. [25], Janel (11T) [26], CRUISE Index [27], and Index26 [27] are all hematological discrimination indices used for discriminating between the IDA and the βTT. However, these indices were obtained empirically and have an inconsistent performance for differential diagnosis of βTT and IDA in the same patient [28]. On the other hand, sometimes, the same indices showed different discrimination power in varied age groups [29, 30].

Recently, the accessibility of powerful statistical software has provided data mining techniques for health-related data. Many studies have proposed advanced statistical methods and data mining techniques such as decision tree methods [31] for differential diagnostic between βTT and IDA to avoid much more expensive, time-consuming, and complicated laboratory procedures and nonsatisfactory hematological indices in discriminating between βTT and IDA [32–38]. [32, 35–39]. Urrechaga, Aguirre, and Izquierdo [39] used multivariable discriminant analysis for differential diagnosis of microcytic anemia. Wongseree et al. [37] implemented neural network and genetic programming for thalassemia classification. Dogan and Turkoglu [35] proposed a decision tree for detecting iron deficiency anemia from hematology parameters.

Jahangiri et al. [32] 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 βTT from IDA. Decision trees have advantages over traditional statistical methods like discriminant analysis and generalized linear models (GLMs). The main advantage of tree-based methods is a tree structure that makes it easy to interpret the clinical data and be accepted by medical researchers and clinicians. CART is one of the best-known classic tree algorithms. However, this algorithm suffers from some problems such as greediness, instability, and bias in split rule selection. Bayesian tree approaches were proposed to solve the greediness of the CART algorithm. The greedy search algorithm has disadvantages such as limit the exploration of tree space, the dependence of future splits to previous splits, generate optimistic error rates, and the inability of the search to find a global optimum [40]. Also, the Bayesian approaches can quantify uncertainty and explore the tree space more than classic tree approaches. Bayesian approaches combine prior information with observations, unlike classic tree methods (these methods use only observations for data analysis). The Bayesian approaches define prior distributions on the components of classic tree methods and then use stochastic search algorithms through Markov Chain Monte Carlo (MCMC) algorithms for exploring tree space [41–47]. So, in the last two decades, many studies have developed Bayesian Treed Generalized Linear Models. These models fit a parametric model such as GLMs instead of using constant models in each tree node. So, these treed algorithms create smaller trees than tree models and improve the tree's interpretation [43].

This paper aims to compare the Bayesian Treed Generalized Linear Models and CART for the differential diagnosis of βTT from IDA based on simple laboratory test results. The outcome variable of the present study is qualitative, so we must use the Bayesian Logit Treed (BLTREED) algorithm for discrimination between these two disorders. This Bayesian treed model fits the logistic regression model in each tree node for data prediction and uses the Metropolis-Hastings algorithm for exploring tree space.

2. Material and Methods

2.1. Criteria for Selecting Patient Groups. In this study, a total of 907 patients aged over 18 years old diagnosed with IDA ($n = 370$) or βTT ($n = 537$) were selected. The mean (±SD) age of the patients was 25 ± 16.1 years. Most of the patients ($n = 592$ (65%)) were women, and 315 (35%) were men.

CBC analysis of EDTA-K2 anticoagulated blood samples was performed using the Sysmex KX-21 automated hematology analyzer (Japan) to measure differential parameters. Hematological parameters like hemoglobin (Hb), 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 for all patients.

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 βTT group, patients had an MCV value below 80 fl. Patients with HbA2 levels of >3.5% were considered as βTT carriers.

2.3. Exclusion Criteria. In the IDA group, the patients who had mutations associated with αTT (3.7, 4.2, 20.5, MED, SEA, THAI, FIL, and Hph) were excluded. For the βTT group, patients with αTT confirmed by mutations in the molecular analysis were excluded. All patients with malignancies or inflammatory/infectious diseases were also excluded.

2.4. Ethical Consideration. This study was approved and supported by the Ethical committee affiliated with the Ahvaz Jundishapur University of Medical Sciences (AJUMS), Ahvaz, Iran. Written informed consent was filled before the enrollment.

2.5. Machine Learning Analysis. Tree-based machine-learning methods are valuable tools in data mining techniques. These methods empower predictive models and could provide a

solution for constructing the diagnostic test with high accuracy [48, 49]. Tree-based models 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 the understanding result of models [50–53]. Tree-based models also were constructed based on Bayesian algorithms. Chipman et al. proposed the Bayesian approach of the CART model (BCART) with defining a prior distribution. Chipman et al. also developed the Bayesian Logit Treed (BLTREED) model as an extension of BCART. The BLTREED model fits a logistic regression model for data prediction in the terminal nodes [43, 54].

2.5.1. Bayesian Logit Treed (BLTREED) Model. The Bayesian approach (BCART) was implemented by using a prior distribution on the two components (Θ, T) of the CART model; T is a binary tree with $\mathcal{K}$ terminal nodes or tree with size $\mathcal{K}$, and $\Theta = (\theta_1, \theta_2, \cdots, \theta_{\mathcal{K}})$ is the parameter set in the terminal nodes ($\theta_i = p_{ij}$, $i = 1, \cdots, \mathcal{K}$, $j = 1, \cdots, N$: the number of distinct classes of the response variable and p_{ij} shows the probability of the jth class of response variable in ith terminal node). The joint posterior distribution of parameters and tree structure was as the following equation:

$$p(\Theta, T) = p(\Theta|T)p(T), \tag{1}$$

where $p(T)$ and $p(\Theta \mid T)$ show the prior distributions for tree and parameters in terminal nodes, respectively.

Usually, the Bayesian approach defines prior distributions as unknown; so, tree structure and parameters in terminal nodes were considered unknown [42]. BCART was extended by fitting a parametric model such as a logistic regression model for data prediction and describing the conditional distribution of $Y|X$ in each terminal node [43, 54]. In the BLTREED model, the conditional distribution of $Y|X$, unlike the BCART model, depends on X $(Y \mid X \sim f(Y \mid X, \theta_i))$ and also by fitting sophisticated model at terminal nodes (by fitting logistic regression model for data prediction in each terminal node), smaller trees and more interpretable were generated. In the BLTREED model, one subset of X can be used to generate the tree and other subsets were used to fit models in terminal nodes (these subsets can be joint and/or disjoint). In the Bayesian approach, $\theta_i = B_i$ shows the regression coefficients for the logistic model fitted in an ith terminal node.

The recursive stochastic process using a tree-generating stochastic process for tree growing $(p(T))$ is as follows [42, 43]:

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

(2) Calculate the probability for splitting node η as follows:

$$P_{\text{Split}} = \alpha\left(1 + d_\eta\right)^{-\beta}, \tag{2}$$

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 that determines the propensity to split decreases with increased tree size.

Actually, α and β are parameters that control the shape and size of trees, and these parameters provide a penalty to avoid an overfitting model

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

The BLTREED model was fitted based on standardized data. So, the same prior distribution can be used independently for parameters in the terminal nodes, and they were considered a multivariate normal distribution with zero mean and variance matrix proportional to the identity for these parameters [43, 54].

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

$$P(T \mid X, y) \propto p(y \mid X, T)p(T). \tag{3}$$

In this study, no informative priors were considered. The priors were uniform on variables at a particular node, and all possible splits for variables.

Where $p(Y \mid X, T)$ is as follows:

$$\begin{aligned} P(Y \mid X, T) &= \int p(y \mid X, \Theta, T)p(\Theta \mid T)\, d\Theta \\ &= \prod_{i=1}^{\mathcal{K}} \int \prod_{h=1}^{n_i} p(y_{ih} \mid x_{ih}, B_i)p(B_i)dB_i, \end{aligned} \tag{4}$$

which $p(y \mid X, \Theta, T)$, (y_{ih}, x_{ih}), and n_i show the data likelihood function, observed values for hth observation in ith node, and the number of observations in ith node, respectively. The integral of equation four has no closed form, so the Laplace approximation was used to solve it [43, 54].

Chipman et al. [42, 43] utilize a Metropolis-Hastings algorithm to simulate equation (3) for finding trees with the high posterior distribution. The Metropolis-Hastings algorithm simulates a Markov chain sequence of trees, namely, $T^0, T^1, T^2, \cdots$.

The simulation algorithm was implemented with multiple restarts for reasons mentioned in Chipman et al. [42, 43].

Table 1: Comparison between hematological parameters of study groups using the Mann–Whitney U test (data are presented as median (IQR)).

	βTT ($n = 537$)	IDA ($n = 370$)	P
MCV (fl)	62 (5.4)	72.2 (9.7)	<0.001
MCH (pg)	19.6 (1.8)	21.9 (4.2)	<0.001
Hb (g/dl)	11 (1.6)	10.5 (2.6)	<0.001
RDW (%)	15.7 (1.7)	15.7 (3.3)	0.94

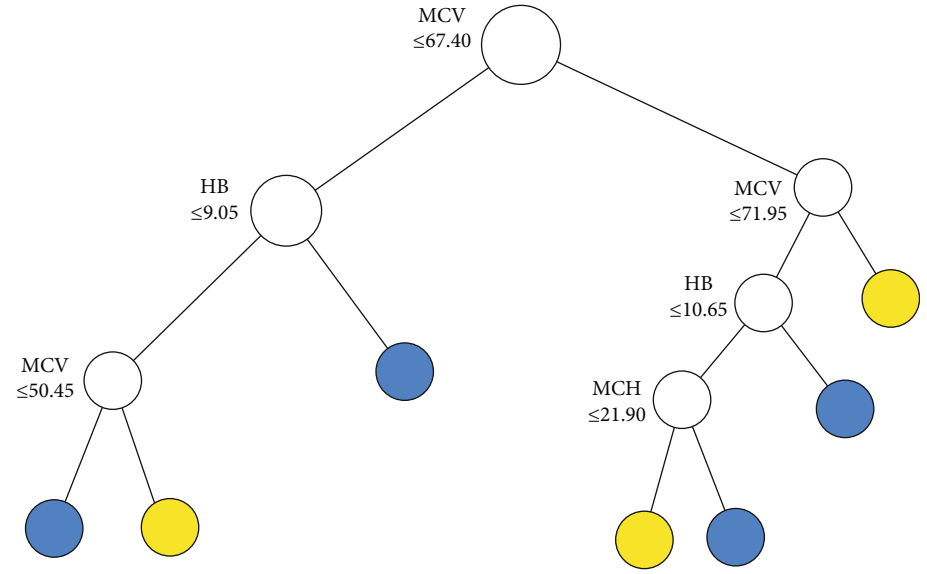

Figure 1: The tree structure of the CART algorithm based on the Gini index (blue terminal node: βTT and yellow terminal node: IDA).

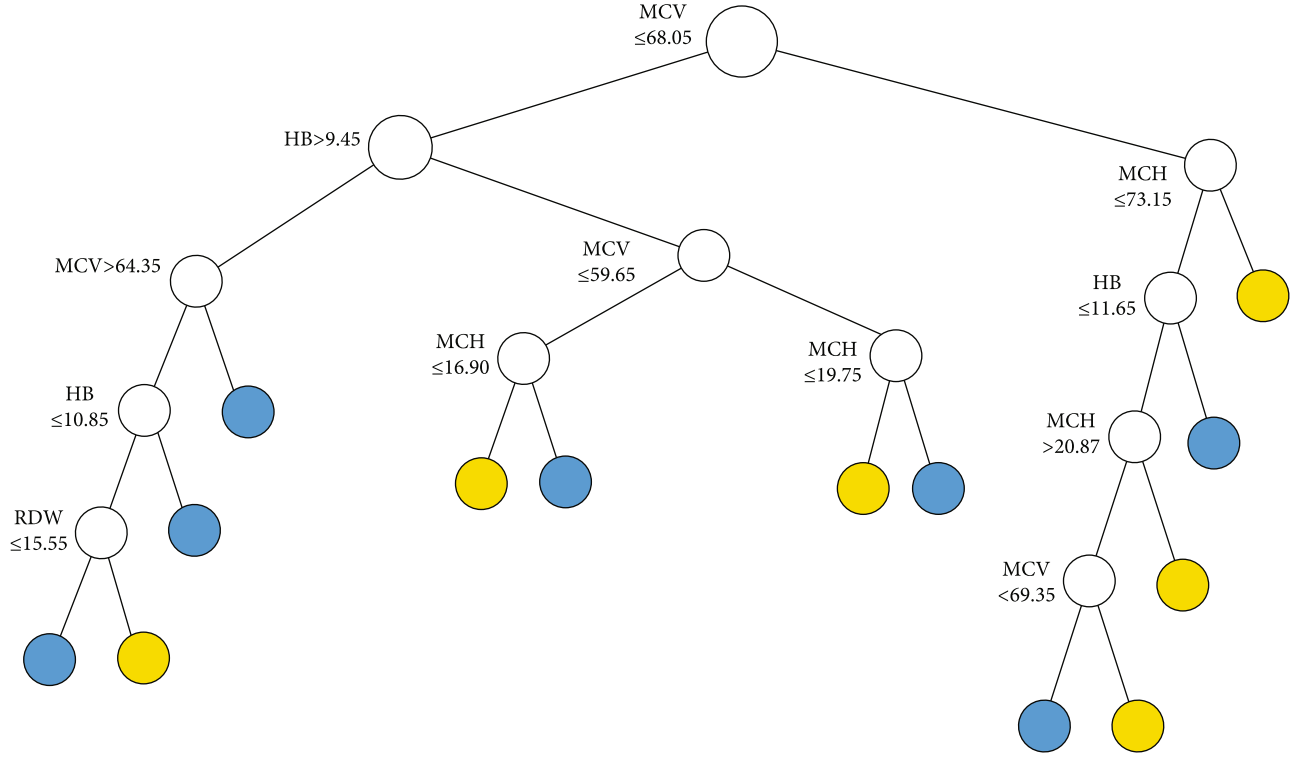

Figure 2: The tree structure of the CART algorithm based on the entropy index (blue terminal node: βTT and yellow terminal node: IDA).

2.5.2. Classification and Regression Trees (CART). Breiman et al. proposed the CART model [55]. The CART algorithm generates a tree using a binary recursive partitioning, and the tree-generating process contains four steps: (1) tree growing: tree growth is based on a greedy search algorithm, and this algorithm generates a tree by sequentially choosing splitting rules. The CART algorithm uses traditional splitting functions for choosing splitting rules (entropy and Gini index). (2) Tree-growing process continues until none of the nodes can split. (3) Tree pruning: this tree algorithm uses the cost-complexity pruning method for tree pruning to avoid overfitting. This pruning method generates a sequence of pruned trees, and each tree in this sequence is an extension of previous trees. (4) Best tree selection: CART uses an

independent test dataset or cross-validation to estimate the prediction error of each tree and then selects the best tree with the lowest estimated prediction error.

2.6. Data Analysis. The BLTREED model and classic CART algorithm based on the two splitting functions like entropy and Gini index (after that, we named the CART method-based Gini index as CART1 and CART method-based entropy as CART2) were fitted by using predictor variables such as hemoglobin (Hb), mean cell volume (MCV), mean cell hemoglobin (MCH), and red cell distribution width (RDW) for differential diagnosis of βTT from IDA.

The BLTREED model fitted using eight restarts with 6000 iterations per restart and a prior standard deviation of 20 for the logit coefficients [54]. For determining the pair of (α, β), the BLTREED model was fitted with two choices, 0.5 and 0.95 for the α parameter, and four choices for β (a range 0.5-2 by step 0.5), then select the pair of (α, β) that generate the best tree with smallest FNR.

Based on the acceptable method of cross-validation in machine learning studies, for assessing the performance of the three models, the dataset was split randomly in the ratio 2 : 1 into a training and a test dataset, respectively, using a stratified random sample to ensure equal allocation of presences and absences (for a classification tree). The model was then fit to the training dataset, and the set of the best trees was determined. For each tree, the posterior predictive distribution was computed for both the training data and the test dataset; this was implemented for each iteration of the BLTREED algorithms, thus incorporating the uncertainty of the model parameters and the data in the evaluation of models. Finally, the predictive performances were calculated based on the confusion matrix of the posterior predictive distribution for both the training and the test dataset [43, 47, 54, 56, 57].

Differential performance of the Bayesian classification tree and CART 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, and the area under the curve (AUCROC). AUCROC represents the degree of separate ability showing how much the machine learning model can distinguish between the classes (IDA and βTT); actually, it is a global measure of diagnostic accuracy. A perfect classification algorithm has an AUCROC = 1. The interpretation of the AUCROC is described as follows: AUCROC > 0.9: excellent differentiation, AUCROC > 0.8: very good differentiation, AUCROC > 0.7: good differentiation, AUCROC > 0.6: sufficient differentiation, AUCROC > 0.5: bad differentiation, and AUCROC < 0.5: classification method is not useful for discriminating between IDA and β TT [58, 59]. Criteria such as Youden's index, accuracy, PLR, NLR (an excellent diagnostic test has NLR < 0.1 and PLR > 10), and AUC take both sensitivity and specificity into consideration, so that can present the performance of the model more accurately than other criteria. In addition, AUC values were compared using DeLong et al. method [60]. A P value < 0.05 was considered a statistically significant difference.

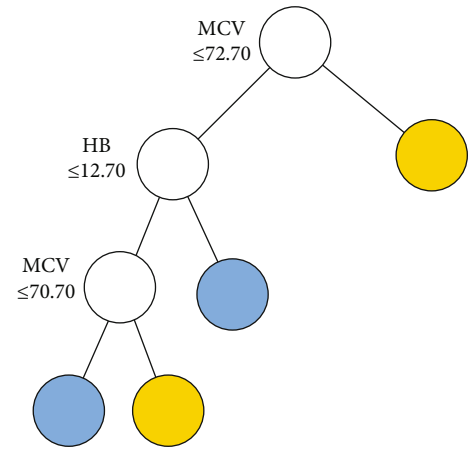

Figure 3: Decision tree for the BLTREED model ($\alpha = 0.95$, $\beta = 1$, Log integrated likelihood = 123.43) (blue terminal node: βTT and yellow terminal node: IDA).

Table 2: Confusion table of the BLTREED model and CART algorithm for training dataset and test dataset.

Dataset	Algorithm	Disease status	TP	FP	FN	TN	(TP+TN)
Training	BLTREED	βTT	363	25	13	234	597
		IDA	234	13	25	363	
	CART1	βTT	366	46	10	213	579
		IDA	213	10	46	366	
	CART2	βTT	358	23	18	236	594
		IDA	236	18	23	358	
Test	BLTREED	βTT	155	8	6	103	258
		IDA	103	6	8	155	
	CART1	βTT	160	33	1	78	238
		IDA	78	1	33	160	
	CART2	βTT	159	12	2	99	258
		IDA	99	2	12	159	

2.7. 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 BLTREED model, R 3.0.3 used for fitting CART algorithm (package rpart), computing performance measures (package ePiR and package pROC), and splitting data to training dataset and test dataset (package caTools).

3. Results

A total of 537 patients were diagnosed as βTT with an average of age (±SD) 22 ± 16.4 including 299 (56%) women and 238 (44%) men, while 370 patients (mean of age (±SD): 29 ± 14.6) were diagnosed as IDA including 293 (79%) women and 77 (21%) men. Table 1 shows the median and interquartile range (IQR) of laboratory parameters as predictor variables across the type of hypochromic microcytic anemia (βTT and IDA).

Table 3: Sensitivity (TPR), specificity (TNR), false-positive rate (FPR), false-negative rate (FNR), positive predictive value (PPV), negative predictive value (NPV), accuracy, Youden's index, positive likelihood ratio (PLR), negative likelihood ratio (NLR), and diagnostic odds ratio (DOR) of the BLTREED model in prediction of IDA and βTT groups and their 95% exact confidence interval for training and test dataset.

	BLTREED		CART1		CART2	
Accuracy measure	Training dataset	Test dataset	Training dataset	Test dataset	Training dataset	Test dataset
TPR	97 (94, 98)	96 (92, 99)	97 (95, 99)	99 (97, 100)	95 (93, 97)	99 (96, 100)
TNR	90 (86,94)	93 (86, 97)	82 (77, 87)	70 (61, 79)	91 (87, 94)	89 (82, 94)
FNR	3 (2, 6)	4 (1, 8)	3 (1, 5)	1 (0, 3)	5 (3, 7)	1 (0, 4)
FPR	10 (6,14)	7 (3, 14)	18 (13, 23)	30 (21, 39)	9 (6, 13)	11 (6, 18)
PPV	94 (91, 96)	95 (91, 98)	89 (85, 92)	83 (77, 88)	94 (91, 96)	93 (88, 96)
NPV	95 (91, 97)	94 (88, 98)	96 (92, 98)	99 (93, 100)	93 (89, 96)	98 (93, 100)
Youden's index	87 (80, 92)	89 (78, 95)	80 (72, 85)	70 (57, 79)	86 (80, 91)	88 (77, 94)
Accuracy	94 (92,96)	95 (91, 97)	91 (89, 93)	87 (83, 91)	93 (91, 95)	95 (91, 97)
PLR	10 (7, 14)	13.36 (7, 26)	5.48 (4, 7)	3.34 (2, 4)	10.72 (7, 16)	9.14 (5, 16)
NLR	0.04 (0.02, 0.07)	0.04 (0.02, 0.09)	0.03 (0.02, 0.06)	0.01 (0, 0.06)	0.05 (0.03, 0.08)	0.01 (0, 0.06)

Table 4: The area under ROC curve (AUC) of BLTREED and CART algorithms in the prediction of IDA and βTT groups for training and test dataset (SE: standard error of AUC; CI: confidence interval).

	BLTREED		CART1		CART2	
	Training dataset	Test dataset	Training dataset	Test dataset	Training dataset	Test dataset
AUC	0.99	0.98	0.93	0.94	0.97	0.97
SE	0.003	0.009	0.011	0.015	0.006	0.011
95% CI	(0.98, 0.99)	(0.96, 0.99)	(0.90, 0.95)	(0.91, 0.97)	(0.96, 0.99)	(0.95, 1)
P value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

The tree structure of CART1, CART2, and BLTREED models is shown in Figures 1–3, respectively. The first split of the three methods of classification trees was based on MCV, which showed that MCV has a higher importance value in differentiation between the βTT and the IDA. Another predictor that was used as the second splitting variable in tree structure was HB. According to the presented trees, the BLTREED model produced a smaller tree size and was more interpretable than the CART algorithm (Figures 1 and 2). This model showed values of MCV ≤ 72.6 screening the βTT patients. The BLTREED model extracted four homogenous subgroups for differentiating between the βTT and the IDA (Figure 3).

The predictive performance of models in differentiation between βTT and IDA was calculated based on the confusion matrix (Table 2). The BLTREED model, CART1, and CART2 trees showed the high TPR, TNR, PPV, NPV, Youden's Index, and accuracy in differentiation between βTT and IDA (Table 3). However, the BLTREED model had a higher accuracy and Youden's index other than CART1 and CART2.

In addition, all the models have NLR < 0.1 that three classification tree algorithms have good diagnostic accuracy for discriminating the patients. Table 4 shows the AUCs of the three tree models from ROC analysis that were statistically significant ($P < 0.001$) and revealed that all three classification methods had an excellent diagnose accuracy (AUC > 0.9: excellent differentiation) in differentiation between the βTT and the IDA. In addition, Figure 4 displays the receiver operating characteristic curves of the BLTREED model, CART1, and CART2 algorithms for the test dataset, and the comparisons of AUC values between the models. According to the exhibited figure, there was no significant difference between the methods ($P > 0.05$).

4. Discussion

In this paper, we used the BLTREED model as the differential diagnostic tool for thalassemia diagnosis. In addition, we compare the predictive performance of the BLTREED model

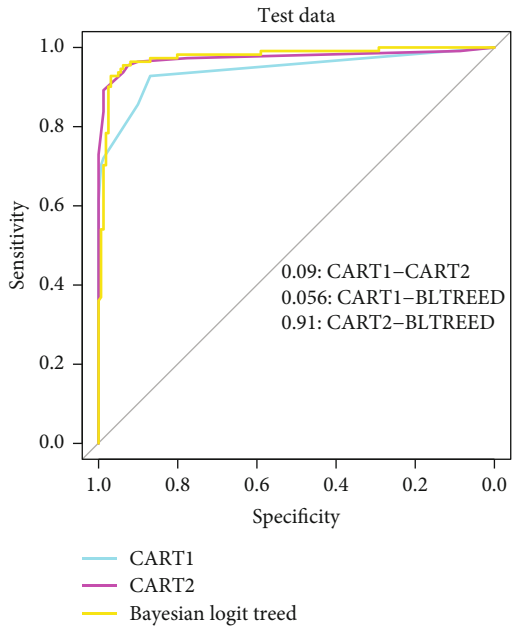

Figure 4: Receiver operating characteristic curves of BLTREED and CART algorithms in the prediction of IDA and βTT groups for test dataset.

as a Bayesian decision tree with the CART algorithm. It is the first study that uses the BLTREED model in the hematological data.

The Bayesian decision tree was used to solve uncertain problems of conventional tree-based methods [43, 54, 61]. This model was implemented by using Hb, MCV, MCH, and RDW as independent variables.

Our dataset included 537 (59%) patients with βTT and 293 (41%) patients with IDA. However, there was not any degree of relative imbalance between the IDA and βTT classes. [62, 63].

Based on our result, MCV and Hb were the main predictor parameters in differential diagnostic, and it showed that the patient with βTT has lower values of MCV.

In previous studies that used the different conventional decision trees for differential diagnosis βTT from IDA, the first split of all algorithms was based on MCV. They also concluded that MCV was a significant predictor variable in the discrimination of IDA and βTT [32, 36]. The performance of the BLTREED model that was evaluated using sensitivity, specificity, false-negative and positive rate, and positive and negative predictive value exhibited the high performance of the differential diagnosis of βTT from IDA. In addition, positive likelihood ratio, negative likelihood ratio, accuracy, and Youden's index showed that BLTREED has good diagnostic accuracy for discriminating the patients. It was indeed classified as 96% of βTT patients. Furthermore, AUC as an overall performance index showed excellent and significant accuracy (99, 98) in training and test data, respectively, in differential diagnostic of βTT and IDA. BLTREED has also generated a tree with a smaller size, and it is more interpretable other than the CART algorithms and indicated better diagnostic performance.

Our study has a limitation, which should be considered. The investigated patients have included just IDA and βTT cases and excluded concomitant diseases and αTT cases. Therefore, considering αTT patients in the study would affect the performance of the presented models and changed the interpretation of the result. Particularly when only simple hematologic parameters are used like in the present study, it may be difficult to distinguish αTT from βTT.

Other studies that used different data mining techniques and decision trees based on the frequentist approach of fitting revealed the high performance and accuracy but lower than our result [32, 34–36, 38]. In many studies which had imbalanced datasets, Oversampling Technique (SMOTE) was applied for handling this problem [34, 64].

The BLTREED model improves the classification performance by solving the uncertainty of previous models [43, 54]. The diagnostic performance of the BLTREED was better than other discrimination methods (classification trees or hematological discrimination indices) in past studies for differentiating βTT from IDA. These studies are as follows: Setsirichok et al. used a C4.5 decision tree, naïve Bayes (NB) classifier, and multilayer perceptron (MLP) for classifying eighteen classes of thalassemia abnormality [38]. Bellinger et al. used classification algorithms like the J48 decision tree, support vector machines (SVM), k-nearest neighbors (k-NN), MLP, and NB for differentiating between βTT, IDA, and cooccurrence of these disorders. In this study, the imbalanced dataset was a cause for the weaker performance [34]. AlAgha et al. compared the diagnostic performance of different classification algorithms such as J48, k-NN, artificial neural networks (ANN), and NB for classifying β-thalassemia carriers. They showed that SMOTE helped decrease the problem of highly imbalanced class distribution and consequently improved the predictive performance [64]. Jahangiri et al. utilized classification tree algorithms such as CHAID, E-CHAID, CART, QUEST, GUIDE, and CRUISE for differential diagnosis of βTT from IDA. They indicated that the CRUISE algorithm has the best diagnostic performance similar to the present study, but this classic algorithm uses the greedy algorithm for tree generating and cannot explore the tree space more than the Bayesian tree approaches. Also, many studies compared the diagnostic performance of hematological discrimination indices, and BLTREED showed better performance in comparison to them [16–19, 23, 25–30, 65–80].

5. Conclusion

In the present study, the BLTREED model showed excellent diagnostic accuracy for differentiating βTT 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 do not need statistical experience. So, it can help physicians in making the right clinical decision.

Abbreviations

βTT: β-Thalassemia trait
IDA: Iron deficiency anemia
MCV: Mean corpuscular volume
MCH: Mean corpuscular hemoglobin
RDW: Red Blood Cell Distribution Width
MCHC: Mean corpuscular hemoglobin concentration
RBC: Red blood cell
BLTREED: Bayesian Logit Treed
TPR: Sensitivity
TNR: Specificity
FNR: False-negative rate
FPR: False-positive rate
NPV: Negative predictive value
PPV: Positive predictive value
PLR: Positive likelihood ratio
NLR: Negative likelihood ratio.

Data Availability

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

Ethical Approval

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

Disclosure

This paper is part of the thesis of Mina Jahangiri, MSc student of Biostatistics (no. U-95095).

Conflicts of Interest

The authors declare that they have no competing interests.

Authors' Contributions

ASM and MJ performed the conception and design, analysis and interpretation of the data, and drafting of the article. FR and NS performed the conception and design, collection and assembly of data, and drafting of the article. All authors approved the final version of the article for submission.

Acknowledgments

This paper was supported by the vice chancellor for Research Affairs of Ahvaz Jundishapur University of Medical Sciences.

References

[1] A. Batebi, A. Pourreza, and R. Esmailian, "Discrimination of beta-thalassemia minor and iron deficiency anemia by screening test for red blood cell indices," *Turkish Journal of Medical Sciences.*, vol. 42, no. 2, pp. 275–280, 2012.

[2] L. Hallberg, "Iron requirements," *Biological Trace Element Research*, vol. 35, no. 1, pp. 25–45, 1992.

[3] W. Mentzer, "Differentiation of iron deficiency from thalassaemia trait," *The Lancet.*, vol. 301, no. 7808, p. 882, 1973.

[4] I. Shine and S. Lal, "A strategy to detect β-thalassaemia minor," *The Lancet.*, vol. 309, no. 8013, pp. 692–694, 1977.

[5] J. England and P. Fraser, "Differentiation of iron deficiency from thalassaemia trait by routine blood-count," *The Lancet.*, vol. 301, no. 7801, pp. 449–452, 1973.

[6] G. G. Klee, V. F. Fairbanks, R. V. Pierre, and M. B. O'sullivan, "Routine erythrocyte measurements in diagnosis of iron-deficiency anemia and thalassemia minor," *American Journal of Clinical Pathology*, vol. 66, no. 5, pp. 870–877, 1976.

[7] P. Srivastava and J. Bevington, "Iron deficiency and/or thalassaemia trait," *The Lancet.*, vol. 301, no. 7807, p. 832, 1973.

[8] B. Ricerca, S. Storti, G. d'Onofrio et al., "Differentiation of iron deficiency from thalassaemia trait: a new approach," *Haematologica*, vol. 72, no. 5, pp. 409–413, 1986.

[9] R. Green and R. King, "A new red cell discriminant incorporating volume dispersion for differentiating iron deficiency anemia from thalassemia minor," *Blood Cells*, vol. 15, no. 3, pp. 481–495, 1989.

[10] J. D. Bessman and D. Feinstein, "Quantitative anisocytosis as a discriminant between iron deficiency and thalassemia minor," *Blood*, vol. 53, no. 2, pp. 288–293, 1979.

[11] A. D. Gupta, C. Hegde, and R. Mistri, "Red cell distribution width as a measure of severity of iron deficiency in iron deficiency anaemia," *The Indian Journal of Medical Research*, vol. 100, pp. 177–183, 1994.

[12] S. Jayabose, J. Giamelli, O. Levondoglu Tugal, C. Sandoval, F. Ozkaynak, and P. Visintainer, "# 262 differentiating iron deficiency anemia from thalassemia minor by using an RDW-based index," *Journal of Pediatric Hematology/Oncology*, vol. 21, no. 4, p. 314, 1999.

[13] O. A. TELMISSANI, S. KHALIL, and G. T. ROBERTS, "Mean density of hemoglobin per liter of blood: a new hematologic parameter with an inherent discriminant function," *Laboratory Hematology*, vol. 5, pp. 149–152, 1999.

[14] A. R. Huber, C. Ottiger, L. Risch, S. Regenass, M. Hergersberg, and R. Herklotz, "Thalassemie-syndrome: klinik und diagnose," Schweiz Med Forum, 2004.

[15] N. KOHAN and M. Ramzi, "Evaluation of sensitivity and specificity of Kerman index I and II in screening beta thalassemia minor," 2008.

[16] M. Sirdah, I. Tarazi, E. Al Najjar, and H. R. Al, "Evaluation of the diagnostic reliability of different RBC indices and formulas in the differentiation of the β-thalassaemia minor from iron deficiency in Palestinian population," *International Journal of Laboratory Hematology*, vol. 30, no. 4, pp. 324–330, 2008.

[17] M. Ehsani, E. Shahgholi, M. Rahiminejad, F. Seighali, and A. Rashidi, "A new index for discrimination between iron deficiency anemia and beta-thalassemia minor: results in 284 patients," *Pakistan journal of biological sciences: PJBS.*, vol. 12, no. 5, pp. 473–475, 2009.

[18] B. Keikhaei, "A new valid formula in differentiating iron deficiency anemia from ß-thalassemia trait," *Pakist J Med Sci.*, vol. 26, pp. 368–373, 2010.

[19] A. A. N. Nishad, A. Pathmeswaran, A. Wickremasinghe, and A. Premawardhena, "The Thal-index with the BTT prediction.exe to discriminate β-thalassaemia traits from other microcytic anaemias," *Thalassemia Reports*, vol. 2, no. 1, 2012.

[20] K. Wongprachum, K. Sanchaisuriya, P. Sanchaisuriya, S. Siridamrongvattana, S. Manpeun, and F. P. Schlep, "Proxy

indicators for identifying iron deficiency among anemic vegetarians in an area prevalent for thalassemia and hemoglobinopathies," *Acta Haematologica*, vol. 127, no. 4, pp. 250–255, 2012.

[21] P. Dharmani, K. Sehgal, T. Dadu, R. Mankeshwar, A. Shaikh, and S. Khodaiji, "Developing a new index and its comparison with other CBC-based indices for screening of beta thalassemia trait in a tertiary care hospital," *International Journal of Laboratory Hematology*, vol. 35, p. 118, 2013.

[22] S. Pornprasert, A. Panya, M. Punyamung, J. Yanola, and C. Kongpan, "Red cell indices and formulas used in differentiation of β-thalassemia trait from iron deficiency in Thai school children," *Hemoglobin*, vol. 38, no. 4, pp. 258–261, 2014.

[23] N. Sirachainan, P. Iamsirirak, P. Charoenkwan et al., "New mathematical formula for differentiating thalassemia trait and iron deficiency anemia in thalassemia prevalent area: a study in healthy school-age children," *Southeast Asian Journal of Tropical Medicine and Public Health*, vol. 45, no. 1, pp. 174–182, 2014.

[24] E. Bordbar, M. Taghipour, and B. E. Zucconi, "Reliability of different RBC indices and formulas in discriminating between β-thalassemia minor and other causes of microcytic hypochromic anemia," *Mediterranean journal of hematology and infectious diseases*, vol. 7, no. 1, 2014.

[25] J. F. Matos, L. Dusse, K. B. Borges, R. L. de Castro, and W. Coura-Vital, "A new index to discriminate between iron deficiency anemia and thalassemia trait," *Revista Brasileira de Hematologia e Hemoterapia*, vol. 38, no. 3, pp. 214–219, 2016.

[26] A. Janel, L. Roszyk, C. Rapatel, G. Mareynat, M. G. Berger, and A. F. Serre-Sapin, "Proposal of a score combining red blood cell indices for early differentiation of beta-thalassemia minor from iron deficiency anemia," *Hematology*, vol. 16, no. 2, pp. 123–127, 2011.

[27] M. Jahangiri, F. Rahim, and A. S. Malehi, "Diagnostic performance of hematological discrimination indices to discriminate between βeta thalassemia trait and iron deficiency anemia and using cluster analysis: introducing two new indices tested in Iranian population," *Scientific Reports*, vol. 9, no. 1, pp. 1–13, 2019.

[28] A. Vehapoglu, G. Ozgurhan, A. D. Demir et al., "Hematological indices for differential diagnosis of beta thalassemia trait and iron deficiency anemia," *Anemia.*, vol. 2014, pp. 1–7, 2014.

[29] F. Rahim and B. Keikhaei, "Better differential diagnosis of iron deficiencyanemia from beta-thalassemia trait," *Turkish Journal of Hematology*, vol. 26, no. 3, pp. 138–145, 2009.

[30] J. J. Hoffmann, E. Urrechaga, and U. Aguirre, "Discriminant indices for distinguishing thalassemia and iron deficiency in patients with microcytic anemia: a meta-analysis," *Clinical Chemistry and Laboratory Medicine (CCLM).*, vol. 53, no. 12, pp. 1883–1894, 2015.

[31] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, no. 2/3, pp. 131–163, 1997.

[32] M. Jahangiri, E. Khodadi, F. Rahim, N. Saki, and A. Saki Malehi, "Decision-tree-based methods for differential diagnosis of β-thalassemia trait from iron deficiency anemia," *Expert Systems*, vol. 34, no. 3, 2017.

[33] M. Maity, T. Mungle, D. Dhane, A. K. Maiti, and C. Chakraborty, "An ensemble rule learning approach for automated morphological classification of erythrocytes," *Journal of Medical Systems*, vol. 41, no. 4, p. 56, 2017.

[34] C. Bellinger, A. Amid, N. Japkowicz, and H. Victor, "Multi-label classification of anemia patients," in *Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on*, IEEE, 2015.

[35] S. Dogan and I. Turkoglu, "Iron-deficiency anemia detection from hematology parameters by using decision trees," *International Journal of Science & Technology.*, vol. 3, no. 1, pp. 85–92, 2008.

[36] E. H. Elshami and A. M. Alhalees, "Automated diagnosis of thalassemia based on data mining classifiers. The International Conference on Informatics and Applications (ICIA 2012)," in The Society of Digital Information and Wireless Communication, 2012.

[37] W. Wongseree, N. Chaiyaratana, K. Vichittumaros, P. Winichagoon, and S. Fucharoen, "Thalassaemia classification by neural networks and genetic programming," *Information Sciences*, vol. 177, no. 3, pp. 771–786, 2007.

[38] D. Setsirichok, T. Piroonratana, W. Wongseree et al., "Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naive Bayes classifier and a multilayer perceptron for thalassaemia screening," *Biomedical Signal Processing and Control*, vol. 7, no. 2, pp. 202–212, 2012.

[39] E. Urrechaga, U. Aguirre, and S. Izquierdo, "Multivariable discriminant analysis for the differential diagnosis of microcytic anemia," *Anemia.*, vol. 2013, pp. 1–6, 2013.

[40] A. S. Malehi and M. Jahangiri, *Classic and Bayesian Tree-Based Methods*, Enhanced Expert Systems, 2019, Intech Open.

[41] D. G. Denison, B. K. Mallick, and A. F. Smith, "A Bayesian CART algorithm," *Biometrika*, vol. 85, no. 2, pp. 363–377, 1998.

[42] H. A. Chipman, E. I. George, and R. E. McCulloch, "Bayesian CART model search," *Journal of the American Statistical Association*, vol. 93, no. 443, pp. 935–948, 1998.

[43] H. Chipman, E. George, and R. McCulloch, "Bayesian treed generalized linear models," *Bayesian statistics.*, vol. 7, pp. 323–349, 2003.

[44] H. A. Chipman, E. I. George, and R. E. McCulloch, "Bayesian treed models," *Machine Learning*, vol. 48, no. 1/3, pp. 299–320, 2002.

[45] Y. Wu, H. Tjelmeland, and M. West, "Bayesian CART: prior specification and posterior simulation," *Journal of Computational and Graphical Statistics*, vol. 16, no. 1, pp. 44–66, 2007.

[46] R. A. O'Leary, J. V. Murray, S. J. Low Choy, and K. L. Mengersen, "Expert elicitation for Bayesian classification trees," *Journal of Applied Probability & Statistics.*, vol. 3, no. 1, pp. 95–106, 2008.

[47] W. Hu, R. A. O'Leary, K. Mengersen, and S. L. Choy, "Bayesian classification and regression trees for predicting incidence of cryptosporidiosis," *PLoS One*, vol. 6, no. 8, article e23903, 2011.

[48] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Chapman and Hall/CRC, New York, 1984.

[49] H. Zhang and B. Singer, *Recursive Partitioning and Applications. Second ed*, P. Bickel, P. Diggle, S. Fienberg, U. Gather, I. Olkin, and S. Zeger, Eds., Springer, New York, 2010.

[50] W. Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.*, vol. 1, no. 1, pp. 14–23, 2011.

[51] G. De'ath and K. E. Fabricius, "Classification and regression trees: a powerful yet simple technique for ecological data analysis," *Ecology*, vol. 81, no. 11, pp. 3178–3192, 2000.

[52] S. C. Lemon, J. Roy, M. A. Clark, P. D. Friedmann, and W. Rakowski, "Classification and regression tree analysis in public health: methodological review and comparison with logistic regression," *Annals of Behavioral Medicine*, vol. 26, no. 3, pp. 172–181, 2003.

[53] N. Speybroeck, D. Berkvens, A. Mfoukou-Ntsakala et al., "Classification trees versus multinomial models in the analysis of urban farming systems in Central Africa," *Agricultural Systems*, vol. 80, no. 2, pp. 133–149, 2004.

[54] W. W. Moe, H. Chipman, E. I. George, and R. E. McCulloch, "A Bayesian treed model of online purchasing behavior using in-store navigational clickstream," *revising for 2nd review at Journal of Marketing Research*, 2002.

[55] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*, CRC Press, 1984.

[56] S. Saha, *Survival Analysis with Bayesian Additive Regression Trees and Its Application*, 2017, https://commons.lib.niu.edu/handle/10843/21175.

[57] H. A. Chipman, E. I. George, and R. E. McCulloch, "BART: Bayesian additive regression trees," *The Annals of Applied Statistics.*, vol. 4, no. 1, pp. 266–298, 2010.

[58] A.-M. Šimundić, "Measures of diagnostic accuracy: basic definitions," *Med Biol Sci.*, vol. 22, no. 4, pp. 61–65, 2008.

[59] L. Donisi, G. Cesarelli, P. Balbi et al., "Positive impact of short-term gait rehabilitation in Parkinson patients: a combined approach based on statistics and machine learning," *Mathematical Biosciences and Engineering*, vol. 18, no. 5, pp. 6995–7009, 2021.

[60] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.

[61] J. B. Gray and G. Fan, "Classification tree analysis using TARGET," *Computational Statistics and Data Analysis*, vol. 52, no. 3, pp. 1362–1372, 2008.

[62] K. Wang, C. A. Phillips, A. M. Saxton, and M. A. Langston, "EntropyExplorer: an R package for computing and comparing differential Shannon entropy, differential coefficient of variation and differential expression," *BMC Research Notes*, vol. 8, no. 1, pp. 1–5, 2015.

[63] Available from: https: //stats. http://stackexchange.com/questions/239973/a-general-measure-of-data-set-imbalance/239982.

[64] A. S. AlAgha, H. Faris, B. H. Hammo, and A.-Z. Ala'M, "Identifying _β_ -thalassemia carriers using a data mining approach: the case of the Gaza Strip, Palestine," *Artificial Intelligence in Medicine*, vol. 88, pp. 70–83, 2018.

[65] J. J. Hoffmann and E. Urrechaga, "Role of RDW in mathematical formulas aiding the differential diagnosis of microcytic anemia," *Scandinavian Journal of Clinical and Laboratory Investigation*, vol. 80, no. 6, pp. 464–469, 2020.

[66] E. Miri-Moghaddam and N. Sargolzaie, "Cut off determination of discrimination indices in differential diagnosis between iron deficiency anemia and β-thalassemia minor," *International journal of hematology-oncology and stem cell research.*, vol. 8, no. 2, pp. 27–32, 2014.

[67] A. Nesa, M. A. Tayab, T. Sultana et al., "RDWI is better discriminant than RDW in differentiation of iron deficiency anaemia and beta thalassaemia trait," *Bangladesh Journal of Child Health.*, vol. 33, no. 3, pp. 100–103, 2010.

[68] C. Beyan, K. Kaptan, and A. Ifran, "Predictive value of discrimination indices in differential diagnosis of iron deficiency anemia and beta-thalassemia trait," *European Journal of Haematology*, vol. 78, no. 6, pp. 524–526, 2007.

[69] M. Ghafouri, S. L. Mostaan, S. Sharifi, G. L. Hosseini, and C. Z. Atar, "Comparison of cell counter indices in differentiation of beta thalassemia minor from iron deficiency anemia," *The Scientific Journal of Iranian Blood Transfusion Organization (KHOON)*, vol. 2, no. 7, pp. 385–389, 2006.

[70] A. Demir, N. Yarali, T. Fisgin, F. Duru, and A. Kara, "Most reliable indices in differentiation between thalassemia trait and iron deficiency anemia," *Pediatrics International*, vol. 44, no. 6, pp. 612–616, 2002.

[71] M. Schoorl, M. Schoorl, J. Linssen 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*, vol. 138, no. 2, pp. 300–304, 2012.

[72] N. Tripathi, J. P. Soni, P. K. Sharma, and M. Verma, "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.*, vol. 2, no. 2, pp. 43–46, 2015.

[73] I. L. Roth, B. Lachover, G. Koren, C. Levin, L. Zalman, and A. Koren, "Detection of β-thalassemia carriers by red cell parameters obtained from automatic counters using mathematical formulas," *Mediterranean journal of hematology and infectious diseases*, vol. 10, no. 1, 2017.

[74] J. F. Matos, L. M. S. A. Dusse, R. V. B. Stubbert et al., "Comparison of discriminative indices for iron deficiency anemia and β thalassemia trait in a Brazilian population," *Hematology*, vol. 18, no. 3, pp. 169–174, 2013.

[75] H. A. Getta, H. A. Yasseen, and H. M. Said, "Hi & Ha, are new indices in differentiation between iron deficiency anemia and beta-thalassaemia trait," *A Study in Sulaimani City-Kurdistan/Iraq IOSR-JDMS*, vol. 14, no. 7, pp. 67–72, 2015.

[76] T. Jameel, M. Baig, I. Ahmed, M. B. Hussain, and M. bin Doghaim Alkhamaly, "Differentiation of beta thalassemia trait from iron deficiency anemia by hematological indices," *Pakistan journal of medical sciences*, vol. 33, no. 3, pp. 665–669, 2017.

[77] L. Tong, J. Kauer, S. Wachsmann-Hogiu, K. Chu, H. Dou, and Z. J. Smith, "A new red cell index and portable rbc analyzer for screening of iron deficiency and thalassemia minor in a chinese population," *Scientific Reports*, vol. 7, no. 1, pp. 1–10, 2017.

[78] C. Shen, "Evaluation of indices in differentiation between iron deficiency anemia and β-thalassemia trait for Chinese children," *Journal of Pediatric Hematology/Oncology*, vol. 32, no. 6, pp. e218–e222, 2010.

[79] E. Urrechaga and J. J. Hoffmann, "Critical appraisal of discriminant formulas for distinguishing thalassemia from iron deficiency in patients with microcytic anemia," Clinical Chemistry and Laboratory Medicine (CCLM), 2017.

[80] M. Jahangiri, F. Rahim, A. Saki Malehi, S. M. S. Pezeshki, and M. Ebrahimi, "Differential diagnosis of microcytic anemia, thalassemia or iron deficiency anemia: a diagnostic test accuracy meta-analysis," *Modern Medical Laboratory Journal.*, vol. 3, no. 1, pp. 1–14, 2019.