scientific reports

OPEN

Machine learning approach for differentiating iron deficiency anemia and thalassemia using random forest and gradient boosting algorithms

Wanicha Tepakhan[1,4], Wisarut Srisintorn[2], Tipparat Penglong[1] & Pirun Saelue[3]✉

Formulas based on red blood cell indices have been used to differentiate between iron deficiency anemia (IDA) and thalassemia (Thal). However, they exhibit varying efficiencies. In this study, we aimed to develop a tool for discriminating between IDA and Thal by using the random forest (RF) and gradient boosting (GB) algorithms. Complete blood count data from 1143 patients with anemia and low mean corpuscular volume were collected (382 patients with IDA, 635 with Thal, and 126 with IDA and Thal). The data were randomly divided into the training and testing datasets in a ratio of 80:20. The RF and GB models had good diagnostic performances for predicting IDA and Thal in the training and testing datasets. In the testing dataset for predicting binary outcomes, GB and RF both had an accuracy of 90.7%, and an area under the receiver operating characteristic curve (AUC-ROC) of 0.953. A lower diagnostic performance was observed when patients with IDA and Thal were included. GB and RF showed accuracies of 80.4% and 82.2%, respectively, and AUC-ROC values of 0.910 and 0.899, respectively. In conclusion, we developed a machine learning approach using GB algorithm. This tool is potentially useful in Thal- and IDA-endemic regions.

Keywords Iron deficiency anemia, Thalassemia, Machine learning, Random forest, Gradient boosting

Anemia is a common condition encountered in clinical practice. It is defined as a low number of red blood cells (RBCs) or a low hemoglobin (Hb) concentration. Anemia is classified into three categories on the basis of the Hb concentration and RBC size: hypochromic microcytic anemia, normochromic normocytic anemia, and macrocytic anemia[1]. Iron deficiency anemia (IDA) and thalassemia (Thal) are the most common causes of hypochromic microcytic anemia. A 2021 global survey reported that the prevalence of anemia was 24.3%, and approximately 66.2% of the total anemia cases are caused by IDA[2]. IDA is characterized by a depleted iron storage that leads to low RBC production. Moreover, it can be caused by a low iron intake, acute or chronic blood loss, or abnormalities in iron absorption. The prevalence of IDA is approximately 1.5–12% in the Thai population[3,4]. Thal, which is one of the most common causes of anemia[2], is an inherited disorder caused by a mutation in the globin gene that results in reduced or absent globin chain production. Patients with Thal traits (TTs) usually exhibit no anemic symptoms. By contrast, patients with Thal disease show widely different clinical phenotypes (from mild to severe anemia) depending on the mutation type. In Thailand, the prevalence rates of TTs, including α-TT, β-TT, and heterozygous Hb E, are approximately 20–30%, 3–9%, and 10–50%, respectively[5].

Laboratory investigations for diagnosing these conditions include serum iron tests, ferritin level assessment, Hb analysis, and deoxyribonucleic acid (DNA) analysis for Thal[6,7]. However, Hb and DNA analyses are unavailable in some hospitals owing to the need for specialized equipment, advanced technical expertise, and the time-consuming nature of the tests. In addition, these investigations can be costly, as patients often incur

[1]Department of Pathology, Faculty of Medicine, Prince of Songkla University, Hat Yai, Songkhla, Thailand. [2]Department of Family Medicine and Preventive Medicine, Faculty of Medicine, Prince of Songkla University, Hat Yai, Songkhla, Thailand. [3]Hematology Unit, Division of Internal Medicine, Faculty of Medicine, Prince of Songkla University, Hat Yai, Songkhla, Thailand. [4]Present address: Centre for Research and Development of Medical Diagnostic Laboratories, Faculty of Associated Medical Sciences, Khon Kaen University, Khon Kaen, Thailand. ✉email: pirun2118@hotmail.com

substantial expenses when physicians request comprehensive confirmatory tests for the diagnosis. Therefore, several mathematical formulas based on RBC indices, including Sirdah, Green and King, Mentzer, England and Fraser, Ehsani, Srivastava, Shine & Lal, and the 11 T score, have been developed to help clinicians select appropriate confirmatory tests for differentiating between IDA and TTs, aiming to reduce investigation costs and time[4,8]. However, the efficiency of these equations varies. Moreover, the cutoff values of each formula are affected by sex, age, and ethnicity, resulting in unsatisfactory sensitivity and specificity results among different populations[9,10]. Applying these formulas for the differential diagnosis between IDA and Thal, including TT and Thal intermedia (TI), offers limited diagnostic value.

In recent years, several machine learning algorithms such as C4.5 decision tree, k-nearest neighbor, artificial neural network, support vector machine, Naive Bayes, random forest (RF), vote algorithm, and extreme learning machine were evaluated for their classification performance in predicting whether a patient has IDA or TT[11–13]. RF showed a high performance with accuracies of 94.2 and 96.0% for IDA and TT, respectively[11,12]. In addition, the gradient boosting (GB) algorithm has been an effective model for predicting and diagnosing several diseases[14]. However, this algorithm has limited information in discriminating between patients with IDA and Thal.

Thus, owing to the advancements of machine learning algorithms and the limitations of previous formulas for the differential diagnosis of IDA and Thal, this study aimed to generate a diagnostic model by using RF and GB algorithms to predict the probability of IDA and Thal. The results of the study should aid clinicians in determining appropriate laboratory investigations for patients with hypochromic microcytic anemia.

Methods

Study population

This cross-sectional study was conducted at Songklanagarind Hospital, the largest tertiary hospital in southern Thailand, between January 2015 and December 2019. We assessed the first-visit data of 7488 patients, who had the following characteristics: (1) age > 15 years, (2) Hb concentration < 13 g/dL in men and menopausal women, or < 12 g/dL in reproductive women, (3) mean corpuscular volume (MCV) < 80 fL, (4) available iron profiles and ferritin level data, and (5) Hb and DNA analyses for Thal. Patients with anemia of inflammation, transfusion-dependent Thal, pregnancy, or incomplete laboratory data were excluded. To exclude anemia due to inflammation and pregnancy, a hematologist reviewed the medical records to confirm the diagnoses of IDA and Thal and to exclude patients with inflammation and infection.

Patients with serum ferritin levels < 30 ng/mL and transferrin saturation < 16% were diagnosed with IDA[15]. All patients were diagnosed with Thal (TT and TI) by using the following diagnostic criteria: patients with Hb type A2A and Hb A2 levels ≥ 3.5% were diagnosed with β-TT. Those with Hb type A2A, Hb A2 levels < 3.5%, and positive α-Thal mutation following DNA analysis were diagnosed with α-TT. Those with Hb type EA and Hb E > 10–35% were considered to have the Hb E trait. Patients diagnosed with TI exhibited Hb patterns such as A2FA, EFA, EE, A2AH, A2ABart'sH, CSA2AH, CSA2ABart'sH, EABart's, EFABart's, CSEABart's, and CSEFABart's; furthermore, these patients had no history of transfusion, and their conditions were confirmed through DNA analysis. The definitions and full names of the abbreviations are shown in the appendix. Patients who met the criteria for IDA and Thal were diagnosed as having IDA with Thal.

Ethical approval

Ethical approval for this study was obtained from the Human Research Ethics Committee (HREC) of the Faculty of Medicine, Prince of Songkla University (REC 62-232-5-2). The HREC waived the requirement for informed consent because this study used deidentified data.

Laboratory techniques

The hematological features were measured using an automated blood cell counter (XN3000; Sysmex Corp., Kobe, Japan). Hb analysis was performed using capillary electrophoresis (CapillaryS2; Sebia, Lisses, France). Serum iron levels, total iron binding capacity, and ferritin levels were measured using an automated analyzer (Cobas e411; Roche, Rotkreuz, Switzerland). DNA analysis for Thal was performed using polymerase chain reaction and reverse dot blot hybridization, as previously described[7,16].

Statistical analysis

The baseline characteristics and hematological features of patients with Thal and IDA were compared using Pearson's chi-squared test for categorical data and the Kruskal–Wallis rank sum test for continuous data. $P < 0.05$ was considered statistically significant. The complete blood count (CBC) data of patients diagnosed with Thal (TT and TI), IDA, and IDA with Thal (TT and TI) were divided into training and testing sets by using a ratio of 80:20. Nine features were used in the machine learning methods, including Hb levels, hematocrit (Hct), MCV, mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), red cell distribution width (RDW), RBC count, age, and sex.

Two types of models were built: binary outcome models (Thal and IDA) and multiclass outcome models (Thal, IDA, and IDA with Thal). Two ensemble machine learning classification methods were used to diagnose Thal and/or IDA. RF builds multiple decision trees on random subsets of the training dataset. This reduces the correlation between trees and avoids overfitting by selecting random subsets of features at each split. The final prediction is the majority vote from all the trees. GB builds trees sequentially by using information learned from previous trees. In theory, this improves accuracy compared to single-decision trees[17].

Our dataset was randomly separated into the training and testing datasets at ratio of 80:20. To minimize over- and under-fitting, the model features in the training dataset were optimized using tenfold cross-validation. The Latin hypercube method was used to sample 1000 sets of feature values for each model. The best feature subsets were those that maximized the AUC-ROC of the validation data. The prediction models were then trained on the entire training data using the best features. To address the class imbalance, the synthetic minority over-sampling technique (SMOTE)[18] was employed to generate synthetic samples for the minority class, thereby effectively balancing the dataset prior to model training.

The performance of the developed models was evaluated on the testing dataset using a range of metrics, including accuracy, Kappa coefficient, sensitivity, specificity and AUC-ROC. Additionally, we compared the diagnostic performances of GB and RF with those of formulas based on RBC indices—such as Hct/Hb, MCV/Hb, Keikhaei, Jayabose, Sirdah, Green and King, Mentzer, England and Fraser, Srivastava, Shine & Lal, Matos, Ricera, Kerman I, Kerman II, Ehsani[8,19–22]—by calculating the difference in AUC-ROC, using the method proposed by Delong et al.[23]. The definitions of all features and formulas are provided in the Appendix. The methodological flowchart of this study is presented in Supplementary Fig. S1.

Analysis was performed using R version 4.4.2[24]. The model specifications and analytical processes were performed using *tidymodels* version 1.2[25]. The underlying analytic packages for RF and GB were *ranger* version 0.17.0 [26] and *xgboost* version 1.7.8.1 [27], respectively. The over-sampling of the minority class was performed using *themis* version 1.0.3[28]. The calculation and comparison of AUC-ROC were performed using pROC version 1.18.5[29].

Results

A total of 14,407 CBC records of patients with anemia and low MCV from 2015 to 2019 were assessed. First-visit data from 7488 patients were selected. In total, 6345 patients with anemia of inflammation (n = 1535), transfusion-dependent Thal (n = 65), or incomplete laboratory records (n = 4745) were excluded. Therefore, 1143 patients were included in the study. The data were randomly divided into two sets namely, the training and testing datasets, by using a ratio of 80:20. Figure 1 shows the distribution of the cases in the training and testing datasets. Table 1 summarizes the baseline characteristics of the study groups. All RBC indices differed significantly among the three groups.

Supplementary Table S1 shows the diagnostic performance of GB and RF for predicting binary outcomes, including Thal and IDA, in the training dataset. In this model, patients with IDA and Thal were not included in the training dataset because the small sample size might have affected the data analysis. The results demonstrated that both GB and RF achieved high accuracy in the training dataset (90.5 and 96.4%, respectively). The AUC-ROC values of GB and RF were 0.969 and 0.996, respectively. However, their performance slightly decreased in the testing dataset, with the accuracy decreasing to 90.7% (95% confidence interval [CI] 86.8–94.6%) for both GB and RF. Table 2 shows that their AUC-ROC remained consistent at 0.953 (95% CI 0.924–0.982).

Supplementary Table S2 shows the diagnostic performance of RF and GB for predicting multiclass outcomes, including Thal, IDA, and IDA with Thal, in the training dataset. RF achieved an accuracy of 91.7% compared with 85.4% for GB. The AUC-ROC values of GB and RF were 0.957 and 0.986, respectively. RF demonstrated a higher sensitivity than GB in predicting Thal and IDA groups. A notable decrease in sensitivity was observed both in RF and GB in predicting IDA with Thal, with RF achieving 87.0% sensitivity and GB achieving 78.0%.

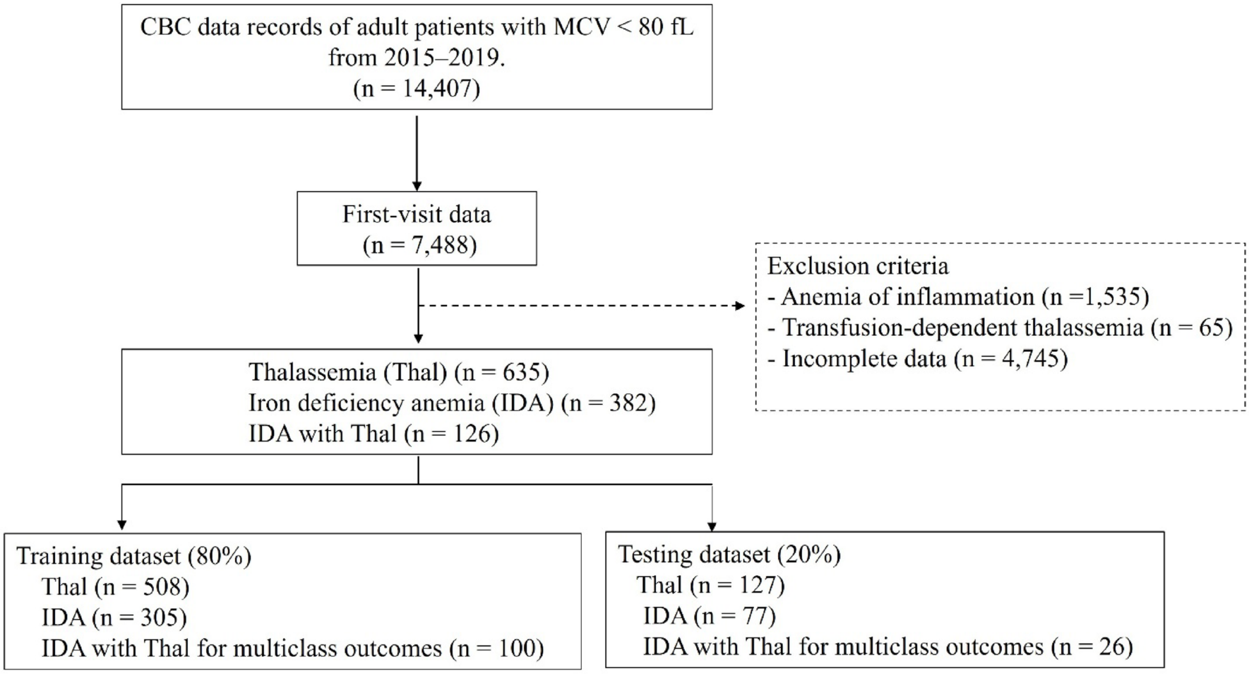
CBC data records of adult patients with MCV < 80 fL
from 2015–2019.
(n = 14,407)
First-visit data
(n = 7,488)
Exclusion criteria
- Anemia of inflammation (n =1,535)
- Transfusion-dependent thalassemia (n = 65)
- Incomplete data (n = 4,745)
Thalassemia (Thal) (n = 635)
Iron deficiency anemia (IDA) (n = 382)
IDA with Thal (n = 126)
Training dataset (80%)
Thal (n = 508)
IDA (n = 305)
IDA with Thal for multiclass outcomes (n = 100)
Testing dataset (20%)
Thal (n = 127)
IDA (n = 77)
IDA with Thal for multiclass outcomes (n = 26)

Fig. 1. Study population flow and distribution of cases in the training and testing dataset. CBC, complete blood count; MCV, mean corpuscular volume

Characteristic	Thal (n = 635) Median (Q1–Q3)	IDA with Thal (n = 126) Median (Q1–Q3)	IDA (n = 382) Median (Q1–Q3)	*P*-value[a]
Sex, female (n, %)	428 (67.4%)	111 (88.1%)	327 (85.6%)	< 0.001
Age (years)	55 (35–69)	43 (28–49)	47 (33–55)	< 0.001
Red blood cells ($\times 10^6$/μL)	4.58 (3.92–5.13)	4.51 (4.17–4.89)	4.31 (3.84–4.71)	< 0.001
Hemoglobin (g/dL)	9.5 (8.3–11.0)	8.8 (7.0–10.3)	8.7 (7.3–9.9)	< 0.001
Hematocrit (%)	30.4 (26.5–34.4)	28.2 (24.2–32.5)	29.6 (26.1–32.9)	< 0.001
Mean cell volume (fL)	68 (62–74)	63 (57–68)	68 (63–74)	< 0.001
Mean corpuscular hemoglobin (pg)	21.1 (19.1–23.9)	19.2 (16.4–21.6)	20.2 (17.6–22.8)	< 0.001
Mean corpuscular hemoglobin concentration (g/dL)	31.9 (30.6–32.9)	30.6 (28.7–31.9)	29.3 (28.2–30.7)	< 0.001
Red blood cell distribution width (%)	17.0 (15.2–21.5)	19.9 (16.7–22.8)	18.7 (16.6–21.2)	< 0.001
Thalassemia type (n)				
α-Thal trait	9	6	0	
Hb Constant Spring trait	18	4	0	
Hb H disease	81	2	0	
Hb H with Constant Spring	19	0	0	
Hb H with Hb E trait	1	0	0	
β-Thal trait	194	26	0	
β^{+}/β^{+}-Thal disease	9	0	0	
β^{+}-Thal/Hb E disease	21	1	0	
Hb E trait	218	80	0	
Hb E trait with Hb Constant Spring trait	2	0	0	
Homozygous Hb E	51	5	0	
HPFH trait	12	2	0	

Table 1. Baseline characteristics and hematological features of patients with Thal and IDA. *Thal* thalassemia, *IDA* iron deficiency anemia, *HPFH* hereditary persistence of fetal hemoglobin. [a]Pearson's chi-squared test; Kruskal–Wallis rank sum test.

	GB		RF	
Metric	Median	95% CI	Median	95% CI
Sensitivity (%)	93.8	89.4–97.7	93.9	89.5–97.7
Specificity (%)	85.7	78.0–92.8	85.9	77.6–93.1
Accuracy (%)	90.7	86.8–94.6	90.7	86.8–94.6
Kappa	0.802	0.719–0.881	0.803	0.717–0.887
AUC-ROC	0.953	0.924–0.982	0.953	0.924–0.982

Table 2. Diagnostic performance of the RF and GB models for predicting binary outcomes (Thal [TT and TI] and IDA) in the testing dataset. Data are shown as median (2.5th–97.5th percentile) from bootstrapping method performed 1000 times. *TT* thalassemia trait, *TI* Thal intermedia, *IDA* iron deficiency anemia, *RF* random forest, *GB* gradient boosting, *AUC-ROC* area under the receiver operating characteristic curve.

Table 3 shows the diagnostic performance of RF and GB for predicting multiclass outcomes, including Thal, IDA, and IDA with Thal, in the testing dataset. Both GB and RF exhibited lower accuracies in the testing dataset (80.4% [95% CI 75.2–85.2%] accuracy for GB and 82.2% [95% CI 77.0–87.0%] for RF) than in the training dataset. The AUC-ROC of both algorithms is also slightly lower than in the training dataset (0.910 [95% CI 0.859–0.949] for GB and 0.899 [95% CI 0.844–0.939] for RF). However, both algorithms maintained high sensitivity for predicting Thal (89.2% [95% CI 83.3–93.9%] for RF and 81.9% [95% CI 74.8–88.6%] for GB). By contrast, the sensitivity (76.8% [95% CI 66.7–86.8%]) was lower in the IDA group using RF. The sensitivity for predicting IDA with Thal was particularly low (69.1% [95% CI 50.0–86.4%] for GB and 65.3% [95% CI 46.4–84.6%] for RF). The two essential variables for predicting multiclass outcomes using GB and RF were MCHC and MCV (Fig. 2).

Table 4 presents the diagnostic performance of 15 previously reported formulas for predicting binary outcomes (Thal and IDA). Among these, only the Hct/Hb index demonstrated strong predictive capability, and it achieved an AUC-ROC of 0.820. Furthermore, our study revealed that the GB and RF algorithms, when utilizing only CBC indices, exhibited significantly higher predictive efficiency than any single index ($P < 0.05$).

Discussion

IDA and Thal are common causes of hypochromic microcytic anemia in Southeast Asia, particularly in Thailand[30]. The differential diagnosis of these abnormalities is vital for effective treatment and proper genetic

Metric	GB			RF		
	Thal	IDA	IDA with Thal	Thal	IDA	IDA with Thal
Sensitivity, % (95% CI)	81.9 (74.8–88.6)	82.1 (73.0–90.4)	69.1 (50.0–86.4)	89.2 (83.3–93.9)	76.8 (66.7–86.8)	65.3 (46.4–84.6)
Specificity, % (95% CI)	88.5 (82.1–94.3)	91.7 (86.8–95.4	90.3 (86.1–94.0)	84.5 (77.0–90.0)	93.6 (89.6–96.9)	92.6 (88.9–96.1)
Accuracy, % (95% CI)	80.4 (75.2–85.2)			82.2 (77.0–87.0)		
Kappa, median (95% CI)	0.669 (0.586–0.751)			0.689 (0.603–0.770)		
AUC-ROC, median (95% CI)	0.910 (0.859–0.949)			0.899 (0.844–0.939)		

Table 3. Diagnostic performance of the RF and GB models for predicting multiclass outcomes (Thal [TT and TI], IDA, and IDA with Thal) in the testing dataset. Data are shown as median (2.5th–97.5th percentile) from bootstrapping method performed 1000 times. *Thal* thalassemia, *TT* Thal trait, *TI* Thal intermedia, *IDA* iron deficiency anemia, *RF* random forest, *GB* gradient boosting, *AUC-ROC* area under the receiver operating characteristic curve.

counseling. Serum ferritin and transferrin saturation are widely used for diagnosing IDA, but their accuracy can be influenced by various confounding factors. For example, serum ferritin thresholds must be adjusted in patients with concurrent inflammation. Moreover, conventional cutoffs for younger adults may not be suitable for older adults because of the cumulative effects of inflammation with age. To enhance accuracy and validity, serum ferritin cutoffs should be adjusted to demographic and physiological factors[31]. Our study used serum ferritin < 30 ng/mL and transferrin saturation < 16% as cutoff values because we included only adult patients without underlying conditions such as inflammation or pregnancy. Several RBC index formulas have been constructed to discriminate between IDA and TT. However, each formula has a different efficiency depending on the study population[19,32,33]. Applying discriminating formulas and indices for TT, TI, and IDA offers limited diagnostic value. Thus, the current study included both TT and TI in the Thal group, which generally occurs in real-world hospital situations. Our internal validation showed that both the RF and GB models performed well in discriminating IDA from Thal in either the training or testing datasets but not in the differential diagnosis of IDA with Thal. Thus, a patient's history review, including data regarding the family history, blood transfusion, history of anemia, blood loss, melena, and hematochezia, might help conduct a proper investigation for the differential diagnosis of IDA with Thal.

In this study, we utilized only RBC indices and personal demographic data to minimize feature redundancy and enhance the performance of our machine learning model. We demonstrated that MCHC and MCV levels are the two important features for machine learning in the RF and GB models, respectively. MCHC represents the average Hb concentration within a single RBC. Notably, this index is significantly lower in patients with IDA than in patients with Thal (TT and TI) (Table 1). This may be explained by the fact that IDA results from a lack of iron, which is essential for Hb production. As a result, RBCs have lower Hb content and are smaller in size. By contrast, Thal is caused by a defect in globin chain production, with iron supply remaining sufficient. Consequently, MCHC is not as drastically reduced in Thal as in IDA. A recent study used GB to predict individuals with TT, IDA, and a normal condition and identified MCV, MCH, RDW-SD, and Hb levels as the most significant features[34]. MCV and MCH are effective markers for screening Thal carriers[35,36]. Similarly, in the current study, MCV also emerged as a key feature in the model, thus aligning with previous findings.

Additionally, we compared the diagnostic performance of previously reported formulas with machine learning models (GB and RF) in binary outcomes model. Among the 15 formulas used to predict Thal (TT and TI) and IDA, only the Hct/Hb index demonstrated strong performance. This is the first study to use the Hct/Hb index to discriminate between IDA and Thal (TT and TI). This index is useful for differential diagnosis because patients with Thal can have low Hb levels[37]. By contrast, patients with IDA have low RBC production, which contributes to low Hb and Hct levels[38]. However, the Hct/Hb index has not been used to differentiate IDA from TT in previous studies. Moreover, the performance of this single index remains lower than that of our machine learning model that uses only RBC indices (Table 4). The remaining 14 formulas demonstrated low performance in our cases because they were originally validated for predicting TT and IDA. However, the Thal group in our study included a diverse range of genotypes that encompasses both TT and TI. Therefore, the applicability of these formulas may be limited in regions where Thal is prevalent.

We demonstrated the efficiency of two predictive models: one distinguishing between Thal and IDA and the other differentiating among Thal, IDA, and the combined group of IDA with Thal by using GB and RF. GB performed better in multiclass outcomes model, with an AUC-ROC of 0.910 (95% CI 0.859–0.949). We suggested the application of a predictive model involving three groups for patient care because we could not exclude patients with both IDA and Thal in a real-world clinical setting. However, the accuracy of GB in the testing dataset was lower than that in the training dataset, and this result might have been due to the small sample size. The model performance of GB in the testing dataset remained high and acceptable. However, further prospective studies should be performed to externally validate the performance and refinement of this diagnostic model.

Finally, a machine learning approach for discriminating between IDA and Thal using the GB algorithm was developed, along with a web-based prediction tool named "PSU Thal-IDA Pred." The probability scores can guide clinicians in selecting suitable confirmation tests in first-visit patients with unknown causes of hypochromic microcytic anemia, resulting in reduced laboratory investigation costs and time and reduced blood volume in

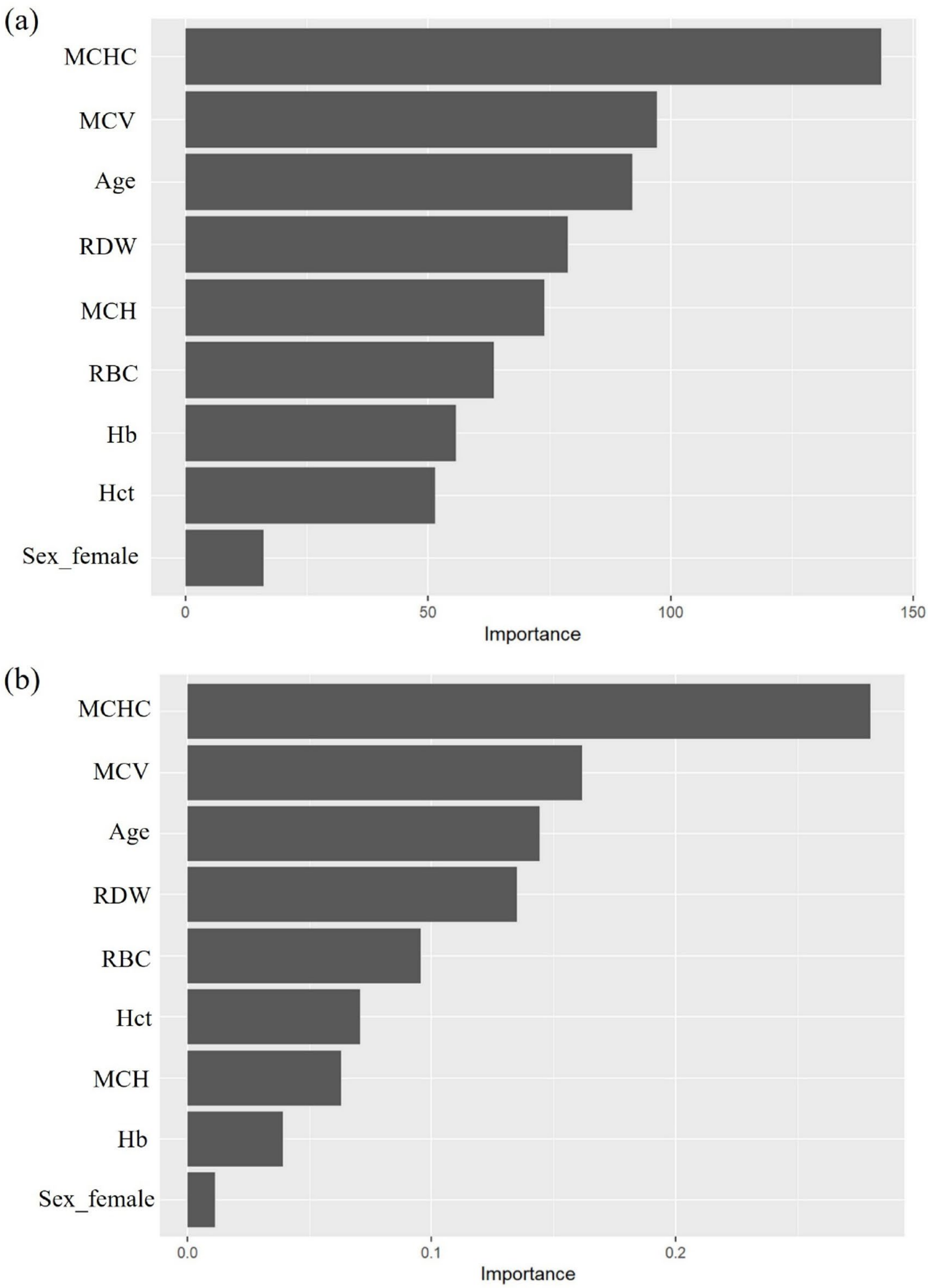

Fig. 2. Variable importance from multiclass outcomes (Thal, IDA, IDA with Thal) of the random forest model (**a**) and gradient boosting model (**b**).

the sample collection of patients with anemia. Users can easily access our website at https://srisintornw.shinyapps.io/small_mcv_prediction_cbc/. The prediction scores were obtained after inputting the RBC indices.

In conclusion, our study demonstrated that machine learning (GB and RF) algorithms are efficient in discriminating between patients with IDA and Thal but not in complex diseases, such as IDA with Thal. Thus, we recommend the application of this diagnostic model for the diagnosis of IDA and Thal in first-visit patients with unknown causes of hypochromic microcytic anemia.

Formula	Cutoff value	95% CI	Sensitivity (%)	95% CI	Specificity (%)	95% CI	Youden's index	95% CI	AUC-ROC of single formula
Mentzer	16.4	13.0–19.7	93.6	93.3–94.0	31.8	27.5–36.2	25.5	21.5–29.6	0.568
Shine & Lal	1221.0	942.5–1499.5	78.4	62.4–94.5	32.0	14.5–49.5	10.5	10.0–11.9	0.529
England and Fraser	9.1	7.3–10.8	88.9	82.0–95.8	44.0	34.6–53.4	32.9	30.4–35.4	0.640
Srivastava	5.1	4.6–5.6	72.5	60.3–84.7	41.4	34.5–48.4	14.0	8.8–19.2	0.478
Green and King	82.7	82.5–82.9	86.7	82.6–90.7	53.0	49.5–56.5	39.6	32.0–47.2	0.664
Jayabose	240.0	232.8–247.2	92.7	92.5–92.9	35.4	30.8–39.9	28.1	23.7–32.5	0.596
Ricera	3.5	3.1–3.9	82.0	68.9–95.1	41.9	25.0–58.8	23.9	20.0–27.7	0.559
Ehsani	21.1	14.5–27.6	70.9	53.3–88.4	54.7	36.7–72.8	25.6	25.0–26.2	0.580
Sirdah	33.0	29.2–36.9	81.6	65.3–97.8	48.8	35.1–62.5	30.4	27.8–33.0	0.643
Kerman_i	381.0	329.7–432.2	79.0	65.2–92.7	35.9	20.3–51.5	14.9	13.0–16.7	0.512
Kerman_ii	94.7	75.9–113.5	73.9	55.2–92.5	51.2	36.2–66.1	25.1	21.4–28.7	0.579
Keikhaei	24.4	24.4–24.4	92.8	92.4–93.2	43.2	41.7–44.7	36.0	34.1–37.9	0.596
Matos	22.8	22.4–23.1	85.6	82.4–88.8	50.0	40.2–59.8	35.6	29.1–42.1	0.686
MCV/Hb	6.5	6.5–6.5	89.8	88.9–90.7	42.6	38.1–47.1	32.4	27.0–37.7	0.635
Hct/Hb	3.2	3.2–3.3	77.9	65.0–90.8	73.8	61.5–86.1	51.7	51.0–52.4	0.820

Table 4. Diagnostic performance of a single formula and comparison of AUC-ROC between each formula with GB and RF algorithms. The AUC-ROC was 0.953 for both GB and RF, and the difference between their AUC-ROC values and those of all single formulas was significant ($P < 0.001$). *GB* gradient boosting, *RF* random forest, *AUC-ROC* area under the receiver operating characteristic curve.

Data availability

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request. Some data may be not available because of privacy or ethical reasons.

Received: 11 December 2024; Accepted: 6 May 2025
Published online: 15 May 2025

References

1. Newhall, D. A., Oliver, R. & Lugthart, S. Anaemia: A disease or symptom. *Neth. J. Med.* **78**, 104–110 (2020).
2. GBD2021 Anaemia Collaborators. Prevalence, years lived with disability, and trends in anaemia burden by severity and cause, 1990–2021: Findings from the Global Burden of Disease Study 2021. *Lancet Haematol.* **10**, e713–e734 (2023).
3. Winichagoon, P. Prevention and control of anemia: Thailand experiences. *J. Nutr.* **132**(Supplement), 862S-866S (2002).
4. Sirachainan, N. 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. J. Trop. Med. Public Health* **45**, 174–182 (2014).
5. Fucharoen, S. & Winichagoon, P. Haemoglobinopathies in Southeast Asia. *Indian J. Med. Res.* **134**, 498–506 (2011).
6. Johnson-Wimbley, T. D. & Graham, D. Y. Diagnosis and management of iron deficiency anemia in the 21st century. *Ther. Adv. Gastroenterol.* **4**, 177–184 (2011).
7. Nopparatana, C., Nopparatana, C., Saechan, V., Karnchanaopas, S. & Srewaradachpisal, K. Prenatal diagnosis of α- and β-thalassemias in southern Thailand. *Int. J. Hematol.* **111**, 284–292 (2020).
8. Pornprasert, S., Thongsat, C. & Panyachadporn, U. Evaluation of applying a combination of red cell indexes and formulas to differentiate β-thalassemia trait from iron deficiency anemia in the Thai population. *Hemoglobin* **41**, 116–119 (2017).
9. Ambayya, A. et al. Haematological reference intervals in a multiethnic population. *PLoS ONE* **9**, e91968 (2014).
10. Huang, T. C. et al. Discrimination index of microcytic anemia in young soldiers: A single institutional analysis. *PLoS ONE* **10**, e0114061 (2015).
11. Laengsri, V. et al. ThalPred: A web-based prediction tool for discriminating thalassemia trait and iron deficiency anemia. *BMC Med. Inform. Decis. Mak.* **19**, 212 (2019).
12. Hasani, M. & Hanani, A. Automated diagnosis of iron deficiency anemia and thalassemia by data mining techniques. *IJCSNS* **17**, 326–331 (2017).
13. Çil, B., Ayyıldız, H. & Tuncer, T. Discrimination of β-thalassemia and iron deficiency anemia through extreme learning machine and regularized extreme learning machine based decision support system. *Med. Hypotheses* **138**, 109611 (2020).
14. Karabayir, I., Goldman, S. M., Pappu, S. & Akbilgic, O. Gradient boosting for Parkinson's disease diagnosis from voice recordings. *BMC Med. Inform. Decis. Mak.* **20**, 228 (2020).
15. Camaschella, C. Iron-deficiency anemia. *N. Engl. J. Med.* **372**, 1832–1843 (2015).
16. Chong, S. S., Boehm, C. D., Higgs, D. R. & Cutting, G. R. Single-tube multiplex-PCR screen for common deletional determinants of alpha-thalassemia. *Blood* **95**, 360–362 (2000).
17. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning: with Applications in R* (Springer US, New York, 2021). https://doi.org/10.32614/CRAN.package.ISLR2.
18. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. Smote: Synthetic minority over-sampling technique. *J. Artifical Intell. Res.* **16**, 321–357 (2002).
19. Pornprasert, S., Panya, A., Punyamung, M., Yanola, J. & Kongpan, C. Red cell indices and formulas used in differentiation of β-thalassemia trait from iron deficiency in Thai school children. *Hemoglobin* **38**, 258–261 (2014).
20. Nalbantoğlu, B. et al. Indices used in differentiation of thalassemia trait from iron deficiency anemia in pediatric population: Are they reliable?. *Pediatr. Hematol. Oncol.* **29**, 472–478 (2012).
21. Demir, A., Yarali, N., Fisgin, T., Duru, F. & Kara, A. Most reliable indices in differentiation between thalassemia trait and iron deficiency anemia. *Pediatr. Int.* **44**, 612–616 (2002).

22. Sirdah, M., Tarazi, I., Al Najjar, E. & Al Haddad, R. Evaluation of the diagnostic reliability of different RBC indices and formulas in the differentiation of the beta-thalassaemia minor from iron deficiency in Palestinian population. *Int. J. Lab. Hematol.* **30**, 324–330 (2008).
23. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44**, 837–845 (1988).
24. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2022). https://www.R-project.org.
25. Kuhn, M. & Wickham, H. *Tidymodels: a Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles* (2020). https://www.tidymodels.org.
26. Wright, M. N. & Ziegler, A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **77**, 1–17 (2017).
27. Chen, T. *et al.* Xgboost: extreme Gradient Boosting. (2022); https://doi.org/10.32614/CRAN.package.xgboost.
28. Themis, H. E. *Extra Recipes Steps for Dealing with Unbalanced* Data. R package. version 1.0.3. (2025). https://doi.org/10.32614/CRAN.package.themis.
29. Robin, X. et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **12**, 77 (2011).
30. Pansuwan, A., Fucharoen, G., Fucharoen, S., Himakhun, B. & Dangwiboon, S. Anemia, iron deficiency and thalassemia among adolescents in Northeast Thailand: Results from two independent surveys. *Acta Haematol.* **125**, 186–192 (2011).
31. Naveed, K. et al. Defining ferritin clinical decision limits to improve diagnosis and treatment of iron deficiency: A modified Delphi study. *Int. J. Lab. Hematol.* **45**, 377–386 (2023).
32. Carla, M. G., Rafael, S. P., Isabel, F. G., Cristina, G. F. & Teresa, S. M. New haematologic score to discriminate beta thalassemia trait from iron deficiency anaemia in a Spanish Mediterranean region. *Clin. Chim. Acta.* **507**, 69–74 (2020).
33. Zaghloul, A. et al. Introduction of new formulas and evaluation of the previous red blood cell indices and formulas in the differentiation between beta thalassemia trait and iron deficiency anemia in the Makkah region. *Hematology* **21**, 351–358 (2016).
34. Wang, W., Ye, R., Tang, B. & Qi, Y. MultiThal-classifier, a machine learning-based multi-class model for thalassemia diagnosis and classification. *Clin. Chim. Acta* **567**, 120025 (2025).
35. Yamsri, S. et al. Prevention of severe thalassemia in northeast Thailand: 16 years of experience at a single university center. *Prenat. Diagn.* **30**, 540–546 (2010).
36. Chaitraiphop, C. et al. Thalassemia screening using different automated blood cell counters: Consideration of appropriate cutoff values. *Clin. Lab.* **62**, 545–552 (2016).
37. Weatherall, D. J. & Clegg, J. B. *The thalassemia syndromes* 4th edn. (Blackwell Science, 2001).
38. Lopez, A., Cacoub, P., Macdougall, I. C. & Peyrin-Biroulet, L. Iron deficiency anaemia. *Lancet* **387**, 907–916 (2016).

Acknowledgements

This study was supported funding by the Faculty of Medicine, Prince of Songkla University (REC 62-232-5-2).

Author contributions

Conceptualization: WT, WS, and PS; Methodology, formal analysis, and investigation: WT, WS, TP, and PS; Writing–original draft preparation: WT and WS; Writing–review and editing: WT, WS, TP, and PS; Funding acquisition: WT.

Declarations

Competing interest

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-01458-5.

Correspondence and requests for materials should be addressed to P.S.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.