International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.3 (2025) pp. 3946-3959
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Machine Learning Classifiers for Differentiation between Iron Deficiency Anaemia and Beta Thalassemia Trait: comparative study

Salma Abdulbaki Mahmood*

University of Basrah, Basrah, Iraq
* **Corresponding Author Email:** Salma.mahmood@uobasrah.edu.iq - **ORCID:** 0000-0002-5247-7851

Article Info:

DOI: 10.22399/ijcesen.2858
Received : 05 April 2025
Accepted : 30 May 2025

Keywords :

Machine Learning,
Microcytosis Anaemia,
Iron Deficiency,
Beta Thalassemia Trait,
Clinical Decision Support

Abstract:

This research compares various machine learning classifiers to differentiate between Iron Deficiency Anaemia (IDA) and Beta Thalassemia trait (BTT). The goal is to identify the most suitable classifiers for handling complex and intertwined medical data, systematically evaluating twelve machine learning (ML) algorithms on a locally curated dataset comprising 2,160 participants (1,080 diagnosed with IDA and 1,040 with BTT conditions). The models being assessed include classical classifiers—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Multilayer Perceptron (MLP), Linear Discriminant Analysis (LDA), k-Nearest Neighbors (KNN), and Naive Bayes—as well as ensemble-based methods: Gradient Boosting, AdaBoost, XGBoost, and CatBoost. Comprehensive pre-processing steps were applied, including outlier data deletion, imputation of missing values, class balancing, and feature scaling. An 80:20 train-test split was employed, with performance validated using 5-fold cross-validation to mitigate overfitting risks. Results indicate that SVM achieved the highest accuracy (97.0%) with an AUC of 99.1%, sensitivity of 95.9%, specificity of 98.7%, and the fastest execution time (0.37 seconds). AdaBoost and MLP followed closely, attaining accuracies of 96.9% and 96.8%, respectively. All models demonstrated high robustness, with F1-scores exceeding 96.8%. SVM provided the best trade-off between diagnostic performance and computational efficiency. These results emphasize the promise of enhanced machine learning models, especially SVM, as dependable and economical diagnostic assistance tools for IDA_BTT differentiation in clinical settings, providing a substitute for costly laboratory tests while maintaining diagnostic precision. The proposed framework underscores the feasibility of deploying ML techniques in haematological diagnostics based on routine clinical data.

1. Introduction

Anaemia is an important global health issue, defined as a condition distinguished by a decrease in the count of red blood cells (Erythrocytes) or a decrease in haemoglobin concentration below normal physiological levels, resulting in impaired Oxygen-carrying capacity and impaired delivery of oxygen to various tissues of the body. According to the latest report by the World Health Organization [1], the global prevalence of anaemia affects approximately 1.6 billion individuals, making it one of the most common and widespread blood disorders globally [2]. Various physiological, environmental, and socio-demographic factors influence anaemia's prevalence and epidemiological distribution, which varies significantly among human populations. These epidemiological patterns are subject to multilevel interactive influences, including demographic characteristics (age and sex distribution), biological factors (racial/ethnic predisposition), high altitude physiological adaptations, health behaviour (smoking exposure patterns), and special physiological conditions such as gestational stage variations. Epidemiological studies indicate that the complex determinant interactions between these determinants lead to significant variations in the epidemiological indicators of anaemia across different populations. This calls for adopting analytical approaches to understand epidemiological patterns and their different dynamics, thus adopting appropriate treatment methods [1].

Iron Deficiency Anemia (IDA) is the most predominant form of anemia, with global epidemiological data indicating that it accounts for approximately 50% of all registered anemia cases globally [4,5]. The World Health Organization (WHO) confirms that this form of anemia is a Major Public Health Concern, especially among vulnerable populations such as women of childbearing age and children under five [1]. IDA is caused by the depletion of the body's iron stores. This results in a disturbance in hemoglobin formation and the appearance of Microcytic Hypochromic Erythrocytes on blood tests. The classic clinical manifestations of this condition are Chronic Fatigue, Cutaneous Pallor, Dyspnea, Palpitations, and Cognitive Impairment. The traditional diagnostic model for IDA relies on specialized laboratory measurements, including Serum Ferritin Levels and the Complete Iron Profile. These expert measurements require a high economic burden, advanced laboratory infrastructure, specialized health and medical personnel, and time delays. These factors pose a significant obstacle in Resource-Limited Settings, necessitating the development of more cost-effective and easy-to-approve alternative diagnostic approaches [2,6].

Beta Thalassemia Trait (BTT) is an inherited hemoglobinopathy that causes microcytosis and mild anemia and tends to manifest masquerading as IDA. HbA2 quantification using HPLC or electrophoresis is required for confirmatory diagnosis, which is not affordable and not feasible in resource-scarce settings [3]. Thalassemia syndromes account for 75% of the documented cases of hemoglobinopathy disorders in Iraq, highlighting a significant public health concern. Recent local epidemiological studies indicate considerable geographic disparities, with Basra province bearing the highest burden, representing 67% of the region's total thalassemia cases. This increased prevalence is primarily linked to the high rate of genetic reasons and consanguineous marriages, facilitating the transmission of recessive hemoglobin disorders through generations. The observed epidemiological trends underscore the urgent need for targeted genetic counseling and comprehensive screening initiatives, particularly in areas with high prevalence, such as southern Iraq [3,4].

Distinguishing between IDA and BTT poses a considerable clinical challenge, mainly because of their shared symptoms, such as fatigue and macrocytosis. Similar laboratory results, including low mean corpuscular volume (MCV) and mean corpuscular hemoglobin (MCH), further complicate the diagnostic process, making it difficult for clinicians to accurately differentiate between these two conditions. Physicians must accurately differentiate between IDA and BTT. Accurate diagnosis is essential to prevent unnecessary iron supplementation and to avoid misdiagnosing major beta thalassemia, particularly during pre-marital consultations aimed at reducing the risk of having children with this condition. This precise distinction safeguards patient health and helps lower healthcare costs associated with inappropriate treatments [7]. Clinical similarities and laboratory overlap in basic blood tests pose a significant diagnostic challenge in differentiating between different types of anemia, including IDA, BTT, and anemia associated with chronic disease (ACD). These similarities lead to severe diagnostic difficulties and risks in developing an appropriate treatment plan, such as in borderline or complex cases (e.g., IDA associated with thalassemia trait).

Given this context, innovative diagnostic systems that utilize artificial intelligence, machine learning, and multivariate data analysis are urgently needed. Such advancements could improve diagnostic accuracy, reduce costs, and expedite access to early and personalized treatment for patients with blood disorders like IDA and BTT, particularly in resource-limited settings [7]. Artificial Intelligence (AI) and Machine Learning (ML) in medical diagnostics, particularly in medical data analysis, show substantial promise. ML algorithms have demonstrated considerable effectiveness in classification, prediction, and image recognition tasks within complex clinical datasets. (Alowais et al., 2023). Despite all of these new developments, there is still no consensus on the best algorithm for differentiation between IDA and BTT based on routine Complete Blood Count (CBC) parameters, which reveals a substantial gap in research in this area.

To address this gap, the current study conducts a comparative analysis of twelve machine learning algorithms, encompassing traditional classifiers and ensemble methods, to evaluate their diagnostic performance in distinguishing IDA from BTT. The aim is to identify the most suitable models for implementation in primary healthcare settings that require adequate accuracy, speed, cost, and scalability strategies.

This study adopts a rigorous six-part structure: (1) an introduction establishing the theoretical framework and research gap; (2) a comprehensive literature review analyzing prior work and methodological limitations; (3) detailed methodology; (4) systematic presentation of analytical results; (5) discussion of scientific interpretation of findings; and (6) evidence-based

conclusions with practical recommendations and future research directions.

2. Literature Review

The following review evaluates the methodologies employed in the few studies that have tackled the differentiation between IDA and BTT, highlighting their findings while identifying limitations and potential areas for enhancement.

Kabootarizadeh et al., Introduced an innovative model that employs artificial neural networks (ANN) to distinguish between iron deficiency anaemia (IDA) and beta-thalassemia (βTT) while also evaluating its effectiveness against conventional diagnostic indicators [8]. The research involved a cohort of 268 patients, 120 with βTT and 148 with IDA, whose complete blood count (CBC) data were sourced from a private laboratory in Iran in 2018. Utilizing MATLAB, several ANN models were constructed, with the selected model achieving remarkable results: a sensitivity of 93.13%, a specificity of 92.33%, and an overall accuracy of 92.5%. This performance surpassed many traditional diagnostic methods, underscoring the potential of artificial intelligence in enhancing medical diagnostics, particularly for blood disorders. Nonetheless, the study encountered methodological limitations, particularly regarding the opacity of ANN decision-making processes, which are often viewed as "black box" systems. This lack of interpretability may hinder their acceptance in clinical environments that prioritize clear diagnostic explanations. Additionally, the model's validation on a larger and more diverse dataset remains untested, raising concerns about its applicability across varied population characteristics and laboratory conditions. Therefore, further research involving more heterogeneous samples is essential to confirm the model's reliability and predictive accuracy in real-world clinical applications.

Laengsri et al., introduced ThalPred, a web-based machine learning tool that employs KNN, DT, RF, ANN, and SVM algorithms to differentiate between β-thalassemia trait (TT) and iron deficiency anemia (IDA) using seven red blood cell parameters [9]. The model, developed using data from 186 Thai patients (146 diagnosed with TT and 40 with IDA), showed remarkable external validation results. It achieved an accuracy of 95.59%, a Matthew's correlation coefficient (MCC) of 0.87, and an AUC of 0.98, outperforming 13 conventional discriminant methods. The research also generated interpretable guidelines from the Random Forest algorithm, which improved clinical transparency. ThalPred is free online (http://codes.bio/thalpred/), simplifying the diagnostic process and making it particularly useful in resource-limited environments. The study does recognize certain limitations, including an imbalance in the sample (with a Thalassemia Trait to Iron Deficiency Anemia ratio of 3.6:1) and the necessity for further validation across various ethnic groups. This suggests that while ThalPred is a useful first-line screening tool, additional research is needed to confirm its effectiveness in more diverse populations.

In the study conducted by Ayyıldız & Arslan Tuncer, a model utilizing machine learning techniques was developed to differentiate between iron deficiency anemia (IDA) and beta-thalassemia (β-thalassemia) [10]. This differentiation relied on red blood cell (RBC) indicators and employed algorithms such as Support Vector Machines (SVM) and K-Nearest Neighbors (KNN). The sample comprised 342 patients, including 152 with thalassemia and 190 with IDA. To identify the most significant features for distinguishing between the two conditions, Neighborhood Component Analysis (NCA) was applied. The optimized model, which combines NCA and SVM, showed impressive results: an accuracy of 95.5%, a sensitivity of 95.8%, and a specificity of 95%. It reached an area under the curve (AUC) score of 97%. These findings emphasize the model's efficiency in differential diagnosis, despite the drawbacks linked to the limited and uneven sample size.

In a study by Xiao et al., a multidimensional analytical model was established to differentiate thalassemia (TT) from iron deficiency anemia (IDA) in pregnant women, utilizing various indicators from red blood cells and reticulocytes [11]. The research involved a sample of 454 anemic pregnant women, with 340 diagnosed with IDA, 66 with α-TT, and 48 with BTT. The authors applied linear discriminant analysis (LDA) to formulate two diagnostic equations: MHA1, aimed at distinguishing IDA from α-TT, and MHA2, designed to differentiate IDA from BTT. These equations incorporated reticulocyte parameters, including the reticulocyte production index and the percentage of small reticulocytes. The results demonstrated that both equations were effective, with MHA1 achieving a sensitivity of 90.9% and specificity of 68.5%, while MHA2 recorded a sensitivity of 81.3% and specificity of 80.3%. Notably, the IDA group exhibited higher MCV, MCH, and MCVm levels than the TT groups. In contrast, HGB and HCT levels were significantly lower in the IDA group relative to α-TT. These findings highlight the importance of reticulocyte parameters in improving diagnostic precision and introducing new equations as economical screening

methods. They potentially reduce reliance on expensive genetic tests, particularly in resource-limited settings. Nonetheless, further validation of these findings in diverse populations remains essential.

Shahmirzalou et al., developed an innovative logistic regression model utilizing routine blood indicators, including red blood cell count (RBC), hemoglobin (HGB), and hemoglobin A_2 (HbA_2), to differentiate between beta thalassemia trait (BTT) and iron deficiency anemia (IDA) in a sample of 292 patients, comprising 73 BTT cases and 219 IDA cases [12]. The findings indicated that HbA_2 was the most distinguishing factor in the differential diagnosis, with each unit increase in HbA_2 correlating to an 8.5-fold increase in the likelihood of IDA (OR = 8.5, $p < 0.001$). The model demonstrated notable diagnostic performance, achieving a sensitivity of 97%, specificity of 72%, and overall accuracy of 93%. However, the study identified methodological weaknesses that could impact the results, such as the imbalance in the distribution of research groups potentially undermining the reliability of the model's validation indicators. Additionally, the exclusion of certain important blood markers for the sake of model simplicity, along with the small sample size, limited the model's completeness and generalizability. The authors cautioned against overinterpreting the model's performance, as it may not accurately reflect its effectiveness in real-world clinical settings, and suggested exploring alternative analytical methods, such as cluster analysis, for a more comprehensive evaluation of the model's diagnostic capabilities.

Saputra et al., introduced an Extreme Learning Machine (ELM) model designed for the multiclass classification of anemia types, including BTT, IDA, HbE, and combined anemias, utilizing complete blood count (CBC) parameters from a cohort of 190 patients at Universitas Gadjah Mada [13]. The model performed exceptionally well, achieving an accuracy of 99.21%. It had a sensitivity of 98.44%, a precision of 99.30%, and an F1-score of 98.84%. These results show that the Extreme Learning Machine (ELM) effectively addresses the issues of traditional diagnostic methods, which often depend on costly confirmatory tests. While offering a rapid and cost-effective solution suitable for resource-limited environments, the study acknowledges the small sample sizes for each anemia type (24 BTT, 41 HbE, 104 IDA, and 21 combined). The research uniquely employs ELM's single-hidden-layer feedforward network architecture in hematological diagnostics, outperforming conventional indices while ensuring computational efficiency. Future research should validate the model with larger, multicenter cohorts to confirm its effectiveness across diverse populations, integrate it into existing clinical workflows, and conduct comparative analyses with deep learning models to evaluate its scalability for rarer anemia subtypes.

In a study conducted by Bahadure et al., the performance of various deep learning models, including Convolutional Neural Networks (CNN), You Only Look Once (YOLO), and Single Shot Multibox Detector (SSD), was evaluated for their effectiveness in a specific application [14]. The models achieved an impressive accuracy rate of 97.6%, based on a dataset comprising 551 samples. Red blood cell (RBC) count, Red Cell Distribution Width (RDW), Mean Corpuscular Volume (MCV), the Mentzer Index, and hemoglobin (Hb) were among the important hematological characteristics that were examined. The findings demonstrated the models' capacity to correctly classify the data by showing substantial statistical differences ($p<0.001$) among the parameters. The study emphasized the significance of these results in a clinical setting and showed that applying cutting-edge machine-learning techniques might significantly increase the diagnostic accuracy of hematological assessments. Better patient outcomes and an improvement in the general standard of clinical treatment could follow from this.

Pullakhandam & McRoy, performed an extensive analysis of NHANES data involving 19,000 participants to assess the effectiveness of machine learning algorithms in identifying iron deficiency anemia (IDA) within populations characterized by class imbalance, where IDA prevalence was noted at 4.9% [15]. The study compared six classifiers utilizing demographic and laboratory data, including complete blood count and serum ferritin, while employing random oversampling techniques to address the minority class imbalance. The gradient boosting (GB) algorithm emerged as the most effective, achieving near-perfect discrimination (AUC=0.99), high accuracy (0.97), and a clinically relevant precision-recall balance (precision=0.82, recall=0.70), surpassing logistic regression and random forests. This research highlights two significant advancements: the effective management of extreme class imbalance through basic oversampling without the need for complex synthetic data generation and the empirical validation of tree-based ensemble methods over traditional linear models for IDA detection in population-level datasets. Although the promising metrics indicate potential for public health screening, the ecological nature of NHANES data necessitates caution in real-world clinical applications, particularly due to the trade-off between high precision and moderate recall, which

may result in missing 30% of actual IDA cases. These results position GB as a strong candidate for anemia surveillance systems. However, they emphasize the importance of prospective validation in clinical environments where prevalence and data quality may significantly differ from those observed in survey conditions.

Uçucu & Azik, introduced an artificial neural network (ANN) model that achieved an impressive 99.5% accuracy in distinguishing between β-thalassemia minor (BTM) and iron deficiency anemia (IDA) by analyzing complete blood count (CBC) parameters from a cohort of 396 patients (216 with IDA and 180 with BTM) [16]. The research highlighted mean corpuscular volume (MCV) and red blood cell (RBC) count as key differentiators, with the Green & King (G&K) and RDWI indices outperforming traditional diagnostic methods ($p<0.001$). While the study confirmed the effectiveness of conventional indices (sensitivity 85-92%, specificity 78-88%), the ANN model significantly enhanced classification accuracy (ΔAUC +0.21 compared to the best traditional index). This AI-driven approach addresses challenges in current diagnostic practices by minimizing dependence on costly confirmatory tests such as hemoglobin electrophoresis and genetic testing, facilitating rapid point-of-care decision-making with processing times under two seconds, and ensuring interpretability through feature importance analysis. The possible clinical use of this model may result in a 40-60% decrease in misdiagnosis rates in primary care environments. Nonetheless, additional validation in various multi-ethnic populations is required because of the study's single-center design.

Al-Najafi et al., developed a population-specific discriminant formula (Karbala formula) using binary logistic regression with stepwise backward elimination to differentiate β-thalassemia trait (BTT) from iron deficiency anemia (IDA) in 1,380 Iraqi adults from Karbala governorate [17]. The novel formula demonstrated superior diagnostic performance (AUC=0.921, 95% CI: 0.905-0.935) compared to 28 established indices, including England-Fraser and MDHL. This study highlights the critical importance of population-specific formula optimization, as the locally derived Karbala formula outperformed conventional indices by 8-12% in diagnostic accuracy ($p<0.001$) within this Middle Eastern cohort. The findings emphasize that hematological discrimination requires regional validation to account for genetic and environmental variations affecting erythrocyte parameters.

Tepakhan et al., created machine learning models, specifically Random Forest (RF) and Gradient Boosting (GB), to distinguish between iron deficiency anemia (IDA) and thalassemia (Thal) using Complete Blood Count (CBC) data from a cohort of 1,143 patients, which included 382 with IDA, 635 with Thal, and 126 with both conditions [18]. The models exhibited strong diagnostic capabilities for binary classification, achieving an accuracy of 90.7% and an AUC-ROC of 0.953. However, their performance was less robust in ternary classification, which accounted for comorbid cases, with accuracy ranging from 80.4% to 82.2% and AUC-ROC values between 0.899 and 0.910. The optimized GB algorithm was subsequently implemented as a web-based tool, "PSU Thal-IDA Pred" (https://srisintornw.shinyapps.io/small_mcv_prediction_v2/), which offers probability scores to assist in confirmatory testing, thereby potentially lowering diagnostic costs and minimizing blood volume requirements in regions where these conditions are prevalent. Table 1 provides an analytical summary of relevant previous research, comparing their methodologies and key findings. Finally, the reviewed studies indicate that a significant methodological challenge faced by most data sets is class imbalance, a prevalent issue in medical applications. This imbalance skews model predictions towards the more dominant class, ultimately compromising the accuracy and reliability of outcomes. Furthermore, no machine learning algorithms have consistently proven to be superior in clinical settings, where high accuracy and dependability are essential for informed diagnostic and therapeutic decisions. Additionally, there is a notable lack of specialized scientific literature addressing this critical issue of discrimination of IDA and BTT, highlighting a knowledge gap that the current research aims to fill.

***Table 1.**: Comparison of discrimination-based ML methods in prior works*

Ref.	Dataset	Size	Target Distribution	Used Parameters	Model Evaluated	Classes	Key Findings	Performance Metrics
[8]	Iran	268	imbalance 148 IDA, 120 βTT	CBC	ANN	IDA vs BTT	ANN	Acc: 92.5% Sens: 93.13% Spec: 92.33%
[9]	Thai adults	186	Imbalance 40 IDA,	CBC	KNN, DT, RF, ANN	IDA vs. TT	SVM outperform	Accuracy: 95.59 %

			146 BTT		and SVM		ed	MCC: 0.87 % AUC: 0.98 %
[10]	Turkish	342	Imbalance 190 IDA, 152 BTT	RBC indices	NCA+SVM, NCA+KNN	IDA vs BTT	NCA+SVM outperformed	Accuracy: 95.5% Precision: 96.7% Sensitivity:95.8% Specificity:95 % F1-score: 96.4%
[11]	China pregnant women	454	imbalance 340 IDA, 66 β-TT, 48 α-TT	Erythrocyte/reticulocyte params	LDA (MHA formulas)	IDA vs BTT/αTT	Developed MHA1 & MHA2 formulas	MHA1: Sens 90.9% MHA2: Spec 80.3%
[12]	Iran	292	imbalance 219 IDA, 73 BTT	RBC, HGB, HbA2	Logistic Regression	IDA vs BTT	HbA2 most significant discriminator	Sens: 97% Spec: 72% Acc: 93%
[13]	Indonesia	190	imbalance 24 BTT, 41HbE, 104 IDA, 21 BTT&IDA, or HbE&IDA	CBC	ELM - SLFNs	IDA vs BTT and HbE	ELM approach has a high performance	accuracy: 99.21%, sensitivity:98.4 %, precision:99.30 %, F1-score: 8.84%.
[14]	India	551	Not specified	CBC	CNN-SSD	Anemia subtypes	AI matched expert diagnosis	Acc: 97.6%
[15]	NHANES	19,000	Balanced by SMOTE 972 IDA 19,203 non-IDA	CBC and ferritin serum	LR, RF, KNN, NB, GB, and GB.	IDA vs non-IDA	Gradient Boost outperformed	Accuracy: 97% Precision: 82% recall :70% AUC: 99%
[16]	Turkish	396	imbalance 216 IDAs 180 BTMs	Hematological variables and blood indices	ANN, DT	IDA vs. BTT	ANN are more powerful	Accuracy: 99.5%
[17]	Iraqi	1380	Imbalance 802 IDA, 578 βTT	Hematologic data + HbA2 Ferritin.	binary LR (stepwise backward elimination)	IDA vs β-TT	New Karbala SCORE	Accuracy: 85.6 % Sensitivity: 85 % Specificity: 86 % AUC: 0.921
[18]	Thail	1,143	382 ID, 635 Thal, 126 IDA and Thal.	RBC indices and demographic data	RF and GB	IDA vs. Thal, IDA vs. IDA+Thal	GB outperformed	Accuracy: 90.7% AUC: 95.3% (for IDA vs. Thal) Accuracy: 80.4% AUC: 91% (IDA, IDA+Thal)

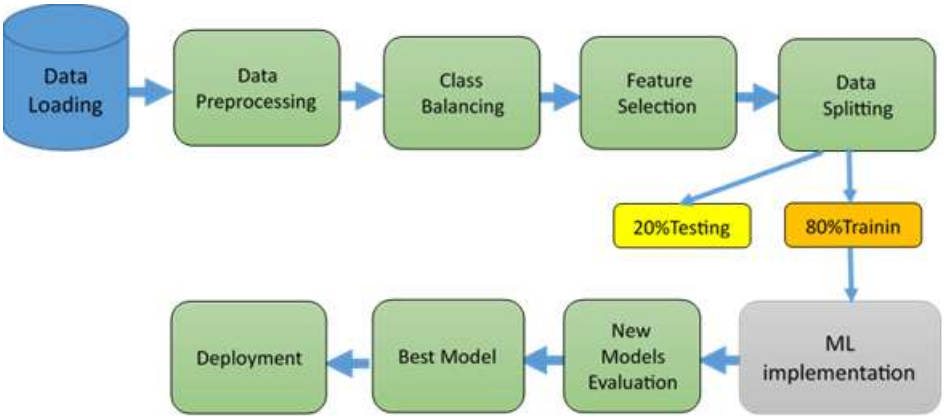

Figure 1. *The framework of the proposed system.*

3. Research methodology

This study employs a rigorous methodological approach encompassing (1) comprehensive data pre-processing including noise reduction and class balancing to ensure data quality; (2) Hyperparameter tuning to optimize model efficiency; (3) stratified 80:20 data partitioning for robust evaluation; (4) implementation of diverse machine learning algorithms with cross validation; and (5) multi-metric performance assessment (accuracy, precision, recall, F1-score, AUC-ROC) to identify optimal models. The selected high-performance model demonstrates exceptional predictive capability, establishing a reliable framework for clinical decision-support systems in haematological diagnostics. Figure 1 shows the general framework of the proposed methodology in this study.

Table 2. *Descriptive Statistics of Hematological Parameters in IDA, BTT*

	IDA		**Thalassemia**	
	Mean	**SD**	**Mean**	**SD**
Hb	9.33	2.92	11.95	4.02
RBC	4.28	0.68	6.06	2.97
MCV	69.64	10.70	64.73	9.71
MCH	21.32	4.92	20.91	4.49
MCHC	30.51	7.76	31.37	2.83

3.1 Data Description

Data for this study were collected from the Basrah Oncology and Hematology Center in Basrah, Iraq, between 2017 and 2020. The study included 2,120 participants: 1,080 individuals diagnosed with IDA (167 males and 913 females), and 1,040 individuals with BTT (569 males and 471 females). Patients with anaemia of inflammation, transfusion-dependent thalassemia, pregnancy, or incomplete laboratory records were excluded. To ensure the exclusion of anaemia resulting from inflammation or pregnancy, a haematologist reviewed the patients' medical records to confirm IDA diagnoses and rule out cases associated with infection or inflammatory conditions. The haematological data presented in this study were meticulously reviewed and validated by physicians and haematologists specializing in anaemia disorders, as acknowledged in the 'Acknowledgments' section.

3.2 Features Distribution

The used dataset contains the following features:
In addition to features, the age class most valued was concentrated between 20 and 60 years. gender was coded as (male =1, female = 0), and IDA was coded as (IDA = 1 for Iron deficiency anemia, and IDA = 0 for BTT).

3.3 Class Balancing

Most medical datasets suffer from the problem of class imbalance, where one class is underrepresented compared to the other. This leads to model bias towards the majority class, poor detection of rare cases, an increased risk of overfitting, and consequently, weakened model performance in diagnosis and evaluation errors. In this study, the dataset exhibits a slight imbalance, consisting of 1080 samples with IDA (the majority class) compared to 1040 BTT(the minority class). To address this issue, the SMOTE (Synthetic Minority Over-sampling Technique) was employed [19]. This technique artificially increases the number of minority class samples while preserving the data's statistical properties and original distribution. As a result, the dataset was balanced to

include 1080 IDA samples and 1080 BTT samples, totaling 2160 samples**.**

3.4 Features Importance and Selection

To enhance performance and accuracy, excluding features that have little impact on the output class is preferable. This can be achieved by using algorithms to calculate feature importance scores, rank them, and then select the most influential features for the final results. The Random Forest method was employed in this study, renowned for its robustness and versatility in machine learning. This method aids in feature selection through its ability to capture complex, non-linear relationships, making it an indispensable tool for optimizing machine learning models and gaining deeper insights into the underlying structure of the data. In our medical dataset, the importance of features such as ['gender', 'Age-class', 'Hb', 'RBC', 'MCV', 'MCH', 'MCHC'] was studied and determined to distinguish IDA and BTT conditions.

3.5 Data Splitting

Splitting the Data into Training and Testing Sets (80% Training, 20% Testing).

3.6 Machine Learning Classifiers Implementation

This study aims to do binary classification (IDA=1 for IDA versus IDA=0 for BTT). We used 12 various ML Methods with default parameters appropriate for binary classification. The following is a Python code for describing the model [20]:

```
models = {
'Logistic Regression': LogisticRegression(random_state=42, max_iter=1000),
'Decision Tree': DecisionTreeClassifier(random_state=42),
'Random Forest': RandomForestClassifier(random_state=42),
'SVM': SVC(probability=True, random_state=42),
'MLPClassifier': MLPClassifier(random_state=42,
max_iter=500),
'LDA': LinearDiscriminantAnalysis(),
'K-Nearest Neighbors': KNeighborsClassifier(),
'Gaussian Naive Bayes': GaussianNB(),
'Gradient Boosting': GradientBoostingClassifier
(random_state=42),
'AdaBoost': AdaBoostClassifier(random_state=42),
'XGBoost': XGBClassifier(random_state=42,
use_label_encoder=False, eval_metric='logloss'),
'CatBoost': CatBoostClassifier(random_state=42, verbose=0)
}
```

3.7 Evaluation Metrics

In this study, we used the following most common metrics used to evaluate the performance of machine learning methods [20], such as,

- Accuracy is the proportion of correctly predicted instances to total instances and measures overall accuracy. The Accuracy can be calculated as follows: $\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$ (1)
- Precision is the proportion of correct optimistic predictions to total predicted optimistic (positives). This measure tells us how well our model avoids false. $\text{Precision} = \frac{TP}{TP + FP}$ (2)
- Recall/ Sensitivity is the ratio of accurate optimistic predictions to the total actual positives, measuring the model's ability to identify all relevant instances.
  $\text{Recall} = \frac{TP}{TP + FN}$ (3)
- F1_Measure is the harmonization of precision and recall, and gives you a balanced measure of

a model's performance.

$$\text{F1\_Measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

- Specificity: The ratio of actual negatives identified adequately by the model.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (5)$$

Where:

  - TP is the number of correctly categorized records.
  - TN is the number of categorized documents that were correctly rejected.
  - FP is the quantity of misclassified records.
  - FN is the percentage of categorized records that have been incorrectly rejected.

- ROC_AUC: ROC, the Receiver Operating Characteristic curve, is a commonly used graphical representation for assessing classification model performance across many threshold specifications. ROC indicates the tradeoff between the True Positive Rate (TPR) and False Positive Rate (FPR) and provides insight into the model's discriminative power. While ROC curves can be useful for assessing binary classifiers and are often used in medical diagnostics, machine learning, and statistical decision theory [21].

4. Results

The computational and software environment utilized for all experiments was a high-performance platform featuring an Intel® Core™ Ultra 9 185H processor, equipped with 24 cores and 28 threads, operating at a base frequency of 2.30 GHz and capable of dynamic boost up to 5.1 GHz. This setup was complemented by 32.0 GB of DDR5 RAM, of which 31.5 GB was available for actual use, all within a Windows 64-bit operating system architecture optimized for x64 processors. The software was developed and implemented using Python version 3.10.12 in the integrated environment Spyder (version 5.4.3), with package management facilitated by Conda (version 23.9.0). This integrated system provided a stable working environment that supports complex data processing and efficient execution of algorithmic models, ensuring optimal compatibility between the advanced processor capabilities and the high-capacity memory allocation, thereby enhancing the reproducibility of results in scientific research.

4.1 Feature Selection Results

The Random Forest algorithm assessed feature importance in the context of the IDA and BTT datasets. The analysis revealed that haemoglobin (Hb) emerged as the most significant predictor with an importance score of 0.37, followed by mean corpuscular haemoglobin concentration (MCHC) at 0.28, mean corpuscular haemoglobin (MCH) at 0.16, mean corpuscular volume (MCV) at 0.1, and red blood cell count (RBC) at 0.07. The variables representing gender and age-class had the least impact, with a score of 0.01. Furthermore, the correlation coefficients indicated strong negative relationships between IDA and the aforementioned haematological parameters: Hb (r = -0.84), MCHC (r = -0.81), MCH (r = -0.76), MCV (r = -0.64), and RBC (r = -0.53), while gender showed a positive correlation (r = 0.42) and Age-class a negative correlation (r =-25), the analysis revealed a positive correlation between gender (r = 0.42) and a negative correlation with age class (r = -0.25). However, these characteristics were excluded from further consideration due to their low significance values. Instead, the focus was narrowed to the more critical hematological parameters, specifically Hb, MCH, MCV, MCHC, and RBC. The analysis results also underscore the crucial role of these haematological indices in diagnosing, with Hb and MCHC being particularly influential. Figure 2 shows the feature importance ranking, which agrees with Spearman's findings in Figure 3.

4.2 ML Models Results

The results in Table 3 demonstrate the clear superiority of ensemble models (including AdaBoost, Gradient Boosting, Random Forest, CatBoost, and XGBoost), as well as MLP Classifier and SVM, in discriminating IDA from BTT. These models achieved classification accuracies exceeding 96%, with notable performance in terms of Precision (≥ 97%), Recall (≥ 95.9%), and AUC (≥ 98.8%). This exceptional performance can be attributed to the ability of ensemble models to integrate the outputs of multiple base learners, enhancing robustness and predictive accuracy. At the same time, MLP Classifier and SVM excelled in capturing non-linear relationships inherent in complex medical datasets. In contrast, traditional models (e.g., Logistic Regression, Decision Tree, Gaussian Naive Bayes, and LDA) showed lower performance, with accuracy not exceeding 94.7% and recall falling below 95.5%. LDA recorded many false negatives (31 cases), indicating its limitations in capturing complex interactions among haematological features (e.g., MCV, MCH, RBC, Hb, and MCHC).

Sensitivity and specificity are key indicators for evaluating the performance of classification models in medical applications, given their direct role in clinical decision support. Sensitivity measures the model's ability to recognize patients with the disease correctly. In contrast, specificity measures the model's ability to identify individuals who do not have the disease correctly. The results of the current study showed that most of the studied models performed well in both metrics, indicating that they can be effectively applied in distinguishing IDA from BTT of the data. In this context, the MLP Classifier model achieved a sensitivity of 96.8% and a specificity of 98.3%, reflecting a high efficiency in detecting infected cases without causing high false alarm rates. Similarly, SVM and CatBoost maintained high levels of sensitivity (above 97%) and specificity (above 97.8%), demonstrating an excellent balance between correct detection and avoiding misdiagnosis. In contrast, the LDA model had the lowest sensitivity (90.1%), indicating a higher probability of missing some cases, which is a concern in sensitive medical contexts that require a high degree of accuracy in early detection. The model with the highest specificity was logistic regression (99.1%), which reduces the likelihood of overdiagnosis that may lead to unnecessary interventions. Considering the time factor is essential when implementing diagnostic systems, particularly when integrated into broader clinical decision support frameworks that demand real-time or near-instantaneous responses. Table 3 above presents each algorithm's training and inference times, enabling an assessment of their computational efficiency. Although ensemble models demonstrated superior accuracy, some, such as CatBoost and MLP Classifier, exhibited longer execution times (13.35 and 18.6 seconds, respectively). In contrast, traditional models like Logistic Regression and Decision Tree achieved significantly faster execution times (0.12 and 0.07 seconds, respectively). While these faster models may be more suitable for time-sensitive applications, their reduced predictive performance necessitates a careful trade-off between computational speed and diagnostic accuracy, depending on the clinical context and system requirements.

The F1-score is one of the pivotal indicators for evaluating the performance of classification models. The harmonic mean between Precision and Recall is an overall metric that reflects a model's ability to minimize false positives and false negatives, and the F1-score balances these two dimensions. In the results of this study, most of the models have high F1-scores, reflecting a good balance in performance. For example, the MLP Classifier model achieved an F1 value of 96.8%, indicating that the model not only achieves high accuracy but also maintains an excellent level of retrieval of infected cases. Similarly, the Random Forest model showed an F1-score of 97.2%, reinforcing its effectiveness in accurate prediction without sacrificing one of its two constituent metrics. On the other hand, models that showed a relatively low F1-score, such as LDA (93.9%), indicate a relative imbalance between precision and recall, which may reflect uneven performance in predicting positive cases, making their use in high-sensitivity medical applications less preferable.

4.3 ML Models ROC AUC Curves

The AUC (Area Under the ROC Curve) is considered a comprehensive and effective metric for evaluating the performance of classification models, as it reflects the balance between the True positive rate (Recall) and the False positive rate (1 - Specificity). The ROC curve illustrates the relationship between these two metrics, and a high AUC value (closer to 1.0) indicates the model's strong ability to distinguish between the target classes (IDA and BTT individuals). Table 3 presents the AUC values for all models. The ensemble models (such as AdaBoost, CatBoost, XGBoost, and Random Forest), along with the SVM model and MLP, achieved near-perfect results (AUC $\geq$ 98%), demonstrating their strong discriminatory power and reinforcing their reliability in supporting critical clinical decisions. In contrast, traditional models recorded AUC values ranging between 95% and 98%, while the Decision Tree model showed the weakest performance, with an AUC of 94.6%, indicating its limited discriminatory capability. Figure 4 depicts the ROC curves for all models utilized in this study. In the lower right corner of the figure, the legend provides the ROC AUC values, with distinct colors assigned to differentiate the performance of each model.

4.4 Comparison with Similar Studies

When comparing the results presented in Table 1 above with previous studies, a clear alignment emerges between our findings and those of the studies [9,10] regarding the superiority of the SVM algorithm, which achieved 95.5% and 95.59% accuracy, respectively. In contrast, our study attained a higher accuracy of 97%, attributed to implementing hyperparameter tuning, cross-validation techniques, and class balancing. Additionally, studies [18] and Pullakhandam &

McRoy, [15] demonstrated the effectiveness of the Gradient Boost algorithm, reporting accuracies of 90.7% and 97%, which aligns closely with our results of 96.7%. Furthermore, studies [8] and [16] highlighted the efficiency of artificial neural network methods in this domain, reinforcing our recommendations regarding the effectiveness of these algorithms in analysing complex data in medical research.

5. Discussion

The first step was an exploratory data analysis (EDA), vital before using machine learning models. This analysis indicated biases in the sample distribution. Using a technique like SMOTE allowed us to address the sample bias. It helped the predictive models perform more consistently with better sensitivity when predicting rare cases and helped ensure some level of fairness in the algorithms across groups. This analysis shows a deep understanding of our data and the ability to rectify biases. These are essential steps toward producing an accurate and equitable artificial intelligence system for medical purposes.

The feature importance analysis using Random Forest provides critical insights into the haematological parameters most discriminant of IDA and BTT. Haemoglobin (Hb) and mean corpuscular haemoglobin concentration (MCHC) emerged as the strongest discriminators alongside MCH, MCV, and RBC. The role of gender and age class was relatively unimportant regarding a feature measure of 0.01. In medical applications, comparative studies on the effectiveness of analytical models represent an essential step in verifying their reliability before they are employed in critical clinical decision-making. Unlike general applications, the medical field demands high levels of accuracy, sensitivity, and performance balance due to the critical consequences of any errors in diagnosis or classification. Consequently, analysing the comparative performance of various machine learning algorithms provides deep insights into each model's relative strengths and weaknesses when handling complex and real-world medical data.

This study applied a range of fundamental traditional classification and ensemble models to distinguish between IDA and BTT samples. When evaluated on essential metrics like accuracy, sensitivity, specificity, and the F1 score, the models showed different performance levels. This variation in results highlights that no one model always does better than another, pointing to the importance of this study in determining the model that best fits the data and the context in which it is applied. Thus, comparative analysis serves as a cornerstone for developing AI-based medical solutions that are both generalizable and clinically reliable. Based on these findings, we recommend using ensemble models, MLP, and SVM classifiers in medical and clinical applications.

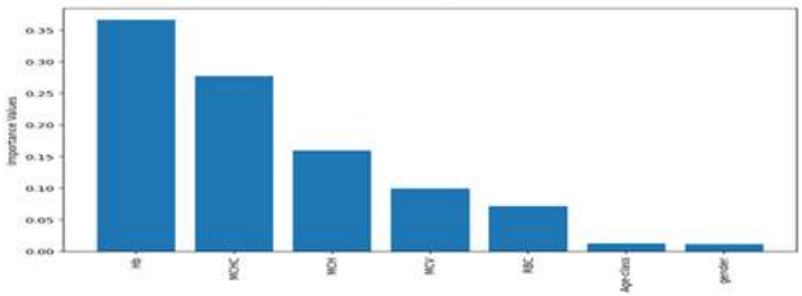

Figure 2. *illustrates the ranking of feature importance.*

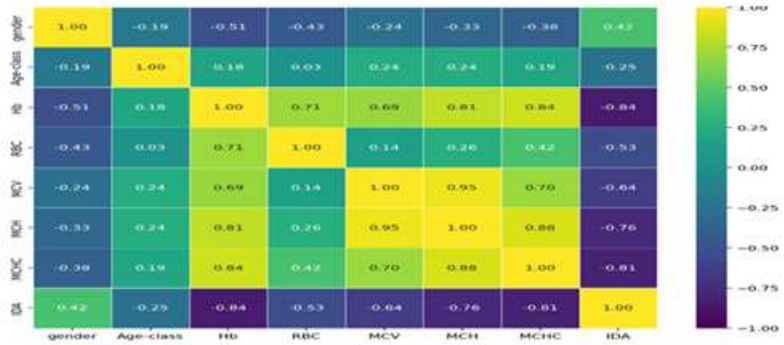

Figure 3. *The Spearman Correlation result for Features importance ranking.*

Table 3. *Performance metrics for machine learning models: Cross-validation, test accuracy, and time spent with the confusion matrix*

ML Models	Confusion Matrix	CV Accuracy	CV Precision	CV Recall Sensitivity	CV F1 Score	CV AUC	specificity	Test Accuracy	Time (s)
MLP Classifier	[[227 4] [9 192]]	96.8	97.8	95.9	96.8	98.8	98.3	97	18.6
SVM	[[228 3] [11 190]]	97	98.2	95.9	97.1	99.1	98.7	96.8	0.37
AdaBoost	[[228 3] [11 190]]	96.9	97.1	96.9	97	99.1	98.7	96.8	0.74
Gradient Boosting	[[226 5] [10 191]]	96.7	97.2	96.4	96.7	99.1	97.8	96.5	1.59
Random Forest	[[225 6] [10 191]]	97.2	97.3	97.2	97.2	99.1	97.4	96.3	1.15
CatBoost	[[226 5] [11 190]]	97	97.4	96.8	97.1	99.2	97.8	96.3	13.35
XGBoost	[[224 7] [11 190]]	97	97.5	96.7	97.1	99	97	95.8	0.55
Decision Tree	[[223 8] [12 189]]	94.6	94.8	94.7	94.7	94.6	96.5	95.4	0.07
K-Nearest Neighbors	[[228 3] [19 182]]	95.7	97.9	93.6	95.7	98	98.7	94.9	0.13
Logistic Regression	[[229 2] [21 180]]	95.1	97.2	93.1	95.1	98.4	99.1	94.7	0.12
Gaussian Naive Bayes	[[227 4] [23 178]]	94.9	96.4	93.5	94.9	98.8	98.3	93.8	0.06
LDA	[[228 3] [31 170]]	94	98	90.1	93.9	98.2	98.7	92.1	0.11

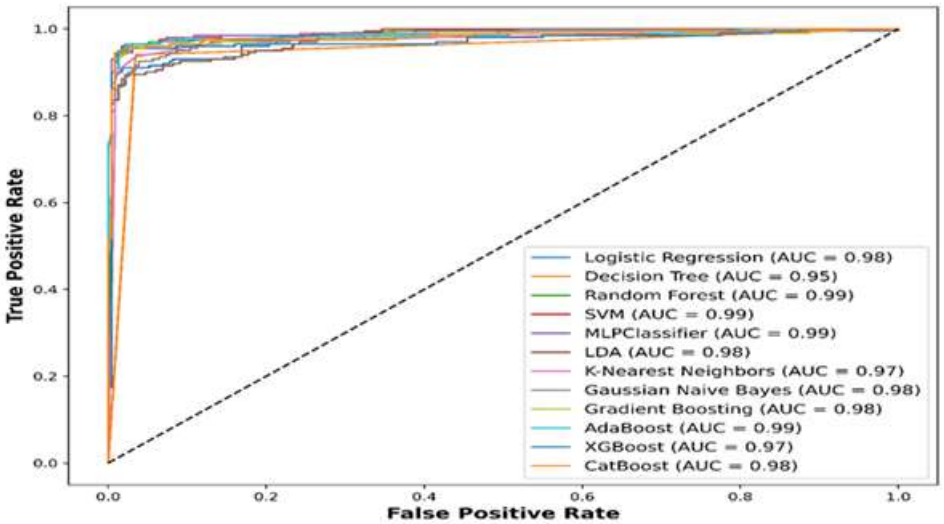

Figure 4. *illustrates the ROC curves of all models employed in this study.*

4. Conclusions

A comprehensive evaluation of twelve machine learning models demonstrated that SVM algorithms (AUC>0.99, execution time 0.37 seconds) and MLP excelled in distinguishing between IDA and BTT samples using fundamental blood indicators. SVM combines high accuracy, speed, and superior discriminative ability, making it suitable for reliable and efficient applications. In light of the results obtained, it is recommended that the SVM model be implemented in real-world medical diagnostic systems. Although the study shows encouraging outcomes, it has notable drawbacks, such as depending on a localized dataset from a single source, using fundamental blood indicators instead of more advanced ones like ferritin and HbA2 variant, and lacking a larger annotated dataset to

increase the generality and reliability. To address these challenges, we propose a future-oriented four-dimensional methodology focusing on external validation across diverse populations and integration with electronic health record systems, incorporating both basic and advanced indicators, developing a hybrid system that combines unsupervised hierarchical clustering, for an automatically annotated dataset with enhanced classification algorithms such as SVM, MLP, CNN, and ensembled ML—finally, implementing model explanation tools (SHAP/LIME) to bolster clinical confidence.

Author Statements:

- **Ethical approval:** The study protocol was also reviewed and approved by the Ethical Committee of the Research Deputy of the University of Basrah, College of Medicine, Basrah, Iraq. Informed consent was obtained from all patients.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** We want to express our sincere gratitude to our esteemed colleagues, Dr. Asaad A. Khalaf, Consultant at Basra Oncology and Haematology Center, and Dr. Saad S. Hamadi, Professor of Internal Medicine, University of Basrah, College of Medicine, for their invaluable contributions to this research. We are also deeply indebted to the Ethical Committee of the Research Deputy at the University of Basrah, College of Medicine, Basrah, Iraq, for their kind cooperation and ethical oversight throughout this study.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1]World Health Organization (Ed.). (2024). Guideline on haemoglobin cutoffs to define anaemia in individuals and populations. *World Health Organization*,

[2]McLean, E., Cogswell, M., Egli, I., Wojdyla, D., & De Benoist, B. (2009). Worldwide prevalence of anaemia, WHO Vitamin and Mineral Nutrition Information System, 1993–2005. *Public Health Nutrition*, 12(04), 444. https://doi.org/10.1017/S1368980008002401

[3]Lafta, R. K. (2023). Burden of Thalassemia in Iraq. *Public Health Open Access*, 7(1), 1–7. https://doi.org/10.23880/phoa-16000242

[4]Khaleed J . Khaleel. (2020). Thalassemia in Iraq Review Article. *Iraqi Journal of Cancer and Medical Genetics*, 13(1), 13–16. https://doi.org/10.29409/ijcmg.v13i1.308

[5]Yang, J., Li, Q., Feng, Y., & Zeng, Y. (2023). Iron Deficiency and Iron Deficiency Anemia: Potential Risk Factors in Bone Loss. *International Journal of Molecular Sciences*, 24(8), 6891. https://doi.org/10.3390/ijms24086891.

[6]Burz, C., Cismaru, A., Pop, V., & Bojan, A. (2019). Iron-Deficiency Anemia. In L. Rodrigo (Ed.), Iron Deficiency Anemia. *IntechOpen.* https://doi.org/10.5772/intechopen.80940

[7]Miri-Moghaddam, E., & Sargolzaie, N. (2014). Cut off Determination of Discrimination Indices in Differential Diagnosis between Iron Deficiency Anemia and β- Thalassemia *Mino*r.8(2):27-32.

[8]Kabootarizadeh, L., Jamshidnezhad, A., Koohmareh, Z., & Ghamchili, A. (2019). Differential Diagnosis of Iron-Deficiency Anemia from beta-Thalassemia Trait Using an Intelligent Model in Comparison with Discriminant Indexes. *Acta Informatica Medica*, 27(2), 78. https://doi.org/10.5455/aim.2019.27.78-84

[9]Laengsri, V., Shoombuatong, W., Adirojananon, W., Nantasenamat, C., Prachayasittikul, V., & Nuchnoi, P. (2019). ThalPred: A web-based prediction tool for discriminating thalassemia trait and iron deficiency anemia. *BMC Medical Informatics and Decision Making*, 19(1), 212. https://doi.org/10.1186/s12911-019-0929-2

[10]Ayyıldız, H., & Arslan Tuncer, S. (2020). Determination of the effect of red blood cell parameters in the discrimination of iron deficiency anemia and beta thalassemia via Neighborhood Component Analysis Feature Selection-Based machine learning. *Chemometrics and Intelligent Laboratory Systems*, 196, 103886. https://doi.org/10.1016/j.chemolab.2019.103886

[11]Xiao, H., Wang, Y., Ye, Y., Yang, C., Wu, X., Wu, X., Zhang, X., Li, T., Xiao, J., Zhuang, L., Qi, H., & Wang, F. (2021). Differential diagnosis of thalassemia and iron deficiency anemia in pregnant women using new formulas from multidimensional analysis of red blood cells. *Annals of Translational Medicine*, 9(2), 141–141. https://doi.org/10.21037/atm-20-7896.

[12]Shahmirzalou, P., Hamze, M. S., & Sadagheyani, H. E. (2024). A New Formula Based on Simple Blood Indices to Differentiate Beta Thalassemia Trait from Iron Deficiency Anemia. *Iranian Journal of Public Health*. https://doi.org/10.18502/ijph.v53i5.15601.

[13]Saputra, D. C. E., Sunat, K., & Ratnaningsih, T. (2023). A New Artificial Intelligence Approach Using Extreme Learning Machine as the Potentially

Effective Model to Predict and Analyze the Diagnosis of Anemia. *Healthcare*, 11(5), 697. https://doi.org/10.3390/healthcare11050697.

[14]Bahadure, N. B., Khomane, R., & Nittala, A. (2024). Anemia detection and classification from blood samples using data analysis and deep learning*. *Automatika*, 65(3), 1163–1176. https://doi.org/10.1080/00051144.2024.2352317

[15]Pullakhandam, S., & McRoy, S. (2024). Classification and Explanation of Iron Deficiency Anemia from Complete Blood Count Data Using Machine Learning. *BioMedInformatics*, 4(1), 661–672. https://doi.org/10.3390/biomedinformatics4010036

[16]Uçucu, S., & Azik, F. (2024). Artificial intelligence-driven diagnosis of β-thalassemia minor & iron deficiency anemia using machine learning models. *Journal of Medical Biochemistry*, 43(1), 11–18. https://doi.org/10.5937/jomb0-38779.

[17] Al-Najafi, W. K., Attiyah, M. N., & Abd, H. M. (2022). Karbala Formula to Differentiate Beta-Thalassemia Trait from Iron Deficiency *Anemia*. 15(1).https://doi.org/10.70863/karbalajm.v15i1.932

[18]Tepakhan, W., Srisint, W., Penglong, T., & Saelue, P. (2025). Machine learning approach for differentiating iron deficiency anemia and thalassemia using random forest and gradient boosting algorithms. *In Review*. https://doi.org/10.21203/rs.3.rs-5623304/v1.

[19]Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357. https://doi.org/10.1613/jair.953

[20]Géron, A. (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow. O'Reilly Media, Inc., *1005 Gravenstein Highway North, Sebastopol, CA* 95472. http://oreilly.com/catalog/errata.csp?isbn=9781491962299

[21]Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010