E-ISSN 2586-9027
Homepage : *https://tci-thaijo.org/index.php/SciTechAsia* **Science & Technology Asia**
Vol.31 No.3 July - September 2026 Page: [381-398]
*Original research article*

# Thalassemia Screening Based on Hematological Data Using Random Forest and Support Vector Machine Models: A Preliminary Study

Mifta Mardiah, Azhari, Rarastoeti Pratiwi, Niken Satuti Nur Handayani*

*Universitas Gadjah Mada, Daerah Istimewa Yogyakarta 55281, Indonesia*

Received 27 December 2025; Received in revised form 24 August 2026
Accepted 11 September 2026; Available online 30 September 2026

**ABSTRACT**

Thalassemia is a hereditary blood disorder caused by impaired globin chain synthesis and represents a public health concern due to increasing incidence and mortality. Screening comprises three levels: hematological examination (Level 1), hemoglobin analysis (Level 2), and molecular genetic testing (Level 3). When Levels 1 and 2 cannot conclusively determine thalassemia status, resource-intensive Level 3 testing is required. This study applied a hybrid exploratory approach combining unsupervised clustering and supervised classification to support Level 1 screening using hematological data from the Thalassemia Research Group. After data preparation, K-means and Gaussian Mixture Model (GMM) clustering were used to identify patterns and outliers, while Random Forest (RF) and Support Vector Machine (SVM) were used for classification. K-means produced six groups, with normal and $\beta$-thalassemia showing relatively clear clustering, although overlap remained. Overlap was also observed between hemoglobin E (HbE) and normal groups and among other categories. GMM achieved 79.11% accuracy. RF achieved the highest classification accuracy of 91%, followed by SVM-linear at 86%. However, SVM-linear correctly classified 40 of 52 test samples, slightly exceeding RF performance. These findings demonstrate the potential of combining unsupervised and supervised learning for Level 1 screening. Larger datasets and additional parameters may improve model performance and generalization.

**Keywords:** Hematological data; Random forest; Screening; Support vector machine; Thalassemia

***Corresponding author:** niken_satuti@ugm.ac.id doi.org/10.70730/sta.v31i3.268410

## 1. Introduction

Thalassemia is a hereditary blood disorder characterized by a decreased level of globin chains, which are fundamental components of hemoglobin [1]. Thalassemia remains a significant global public health concern due to its increasing prevalence and significant clinical burden [2]. The World Health Organization estimates that 6–10% of Indonesia's population are carriers of thalassemia [3]; however, this Fig. is widely considered an underestimate due to limited national screening coverage and reporting limitations [4]. Indonesia's substantial ethnic diversity and high population mobility have contributed to significant genetic heterogeneity, leading to a diversity of thalassemia mutations across various subpopulations [5, 6]. These conditions collectively emphasize the necessity of developing effective, scalable, and accessible screening methods.

A hierarchical diagnostic framework is employed for thalassemia diagnosis, consisting of hematological examination (level 1), hemoglobin analysis (level 2), and molecular genetic testing (level 3). Level 1 screening is based on hematological assessment, including a complete blood count (CBC) and peripheral blood smear evaluation to detect abnormalities in erythrocyte morphology. A mean corpuscular volume (MCV) of less than 80 fL and mean corpuscular hemoglobin (MCH) of less than 27 pg is indicative of clinical suspicion of thalassemia [7-9]. The level 2 assessment utilizes high-performance liquid chromatography (HPLC) or hemoglobin electrophoresis to quantitatively identify hemoglobin variations [7]. Previous studies have demonstrated that HPLC-based screening is effective in detecting various thalassemia phenotypes, including $\beta$-thalassemia, $\alpha$-thalassemia carriers, $\alpha\beta$-thalassemia, and hemoglobin E (HbE carriers) [10].

Furthermore, when hematological and hemoglobin investigations yield inclonclusive results, molecular analysis (level 3) provides a definitive diagnosis; however, it is resource-intensive, time-consuming, and has limited accessibility in various regions, thereby limiting its efficacy as an essential screening tool [7, 8].

In this context, artificial intelligence (AI) offers a practical approach to enhancing early-stage thalassemia screening. Level 1 AI-powered analysis facilitates a clearer interpretation of hematological data and reduces the subsequent reliance on molecular testing. AI systems can analyze complex patterns in hematological parameters to assist in computer-aided screening for thalassemia prior to genetic testing. This approach is particularly beneficial in environments with limited laboratory resources and specialized expertise.

AI, characterized as computational systems capable of learning, reasoning, and executing tasks traditionally requiring human intelligence [11], has progressively been applied in biomedical research for the analysis of complex, high-dimensional datasets. AI-powered pattern recognition improves diagnostic accuracy in clinical evaluations, facilitates early intervention, and reduces healthcare costs [12].

Supervised algorithms such as Random Forest (RF) and Support Vector Machine (SVM) are widely used for classification tasks in machine learning. The RF approach achieves improved predictive accuracy by aggregating multiple decision trees [13]. On the other hand, SVMs identify optimal hyperplanes for capably separating data classes, particularly in high-dimensional spaces [14].

In addition, unsupervised learning strategies offer additional benefits by re-

vealing hidden patterns within unlabelled data. Clustering strategies such as K-means and Gaussian Mixture Models (GMM) enable analysis in distinguishing data clusters, patterns, and potential anomalies [15]. K-means clusters data points based on proximity to centroids [16], while GMMs utilize a probabilistic approach to demonstrate data distribution and assign membership to clusters [17].

The significance of applying AI in thalassemia screening is underscored by the development of predictive tools that can mitigate the need for resource-intensive follow-up examinations [18]. This approach supports the equitable distribution of thalassemia screening in regions of Indonesia with limited access to advanced healthcare services and facilities. The integration of diverse strategies enhances the in-depth understanding of hematological data characteristics prior to classification. A hybrid exploratory approach, combining unsupervised clustering with supervised classification, has been employed to improve Level 1 thalassemia screening through hematological data. Notably, the present study focuses on evaluating the performance metrics of the resulting model, whereas previous studies were restricted to assessing the efficacy of thalassemia screening methodologies. This study demonstrates that AI-driven screening, employing RF and SVM models, offers a more concise, effective, and scalable approach for the early diagnosis of thalassemia using hematological data through multidisciplinary collaboration.

## 2. Materials and Methods

### 2.1 Data

Hematological data were obtained from the Thalassemia Research Group. The study was conducted from August to November 2025. This study was approved by the Research Ethics Commission, Ethical Clearance number KE/FK/1823/EC/2025. The dataset for model construction comprised six categories, $\alpha$-thalassemia (37), $\beta$-thalassemia (55), Iron Deficiency Anemia (IDA) (11), HbE (42), normal (311), and thalassemia major (11) (Table 1). Hematological parameters included hemoglobin (Hb) (g/dL), leukocytes (103/$\mu$L), erythrocytes (106/$\mu$L), platelets (103/$\mu$L), hematocrit (%), MCV (fL), MCH (pg), and Mean Corpuscular Hemoglobin Concentration (MCHC) (g/dL).

**Table 1.** Hematological data used.

| No. | Categories | Number of Data | Number of Training Data | Number of Test Data |
|---|---|---|---|---|
| 1 | IDA | 11 | 9 | 2 |
| 2 | $\alpha$-thalassemia | 37 | 27 | 10 |
| 3 | $\beta$-thalassemia | 55 | 45 | 10 |
| 4 | HbE | 42 | 32 | 10 |
| 5 | Normal | 311 | 301 | 10 |
| 6 | Thalassemia major | 11 | 11 | 10 |
| | Total | 467 | 425 | 52 |

Test data were derived from the complete dataset of each category, except for IDA and thalassemia major categories, due to insufficient data volume.

The dataset was grouped based on HPLC test results and diagnostic notes recorded in the archived sample data of patients or screening participants. In this study, no direct clinical or molecular testing was performed on the samples. Modeling was conducted using available archived data, which was grouped based on the conclusions of the HPLC test results and the notes included in the dataset.

The $\alpha$-thalassemia group lacked information regarding the a0 and a+ subgroups. Therefore, the data were combined into a single $\alpha$-thalassemia category. The $\beta$-thalassemia group consisted of sam-

ples classified as $\beta$-thalassemia in the data archives, whereas the thalassemia major group consisted of samples recorded in the archives as both heterozygous for $\beta$-thalassemia and HbE. The thalassemia major group consisted of 7 samples that have been confirmed through molecular testing. However, the available archived data does not include information regarding the type or number of mutations identified in each sample. Therefore, the grouping used in this study was based on the classifications available in the archived data rather than on additional clinical or molecular confirmation performed in this study.

### 2.2 Data preprocessing

Hematological data were processed into datasets and uploaded to Google Colab [19]. Data preprocessing consisted of several stages, as follows:

a. Data cleaning: checking and deleting null and duplicate data, and converting text data into numerical data [19, 20].

b. Exploratory Data Analysis: data was processed by visualizing it into histograms, boxplots, and correlation graphs [19, 21].

c. Feature selection: features which had the most significant effect on the thalassemia dataset were chosen [22]. Hematologic features considering physiological relevance in the application of machine learning were integrated, to provide a clinical picture in differentiating categories.

d. Data splitting: data was divided into two types of data, namely training data (80%) and test data (20%) [23]. This division of data maintains the proportions of each class so that both subsets remain representative of their original distribution [24].

e. Resampling: a combination of Synthetic Minority Over-sampling Technique (SMOTE) and undersampling to overcome imbalanced data distribution. SMOTE performs the oversampling process by producing synthetic data on minority classes, while undersampling will equalize the amount of data for minority and majority classes or only reduce the amount of data for the majority class [25].

f. Standardization: the scale of data was changed to be uniform so that it could contribute equally to the analysis process [26].

### 2.3 Data processing

#### *2.3.1 Clustering*

Clustering was done using K-Means Clustering and GMM. K-means grouped the data into multiple optimal clusters based on data similarity [27]. Meanwhile, GMM clustered data based on probability distribution [26]. The application of the clustering method helped to uncover the patterns in hematological data that overlap between clinical categories. The clustering results can provide an overview of possible difficulties in classification due to the similarity of hematological characteristics at the supervised learning stage.

#### *2.3.2 Modelling*

Training and testing models was done using RF and SVM to predict the value of the test data to draw final conclusions [23, 28].

The RF model builds many decision trees as a bootstrap sampling, then the prediction results of the entire tree are combined and the final prediction on the classi-

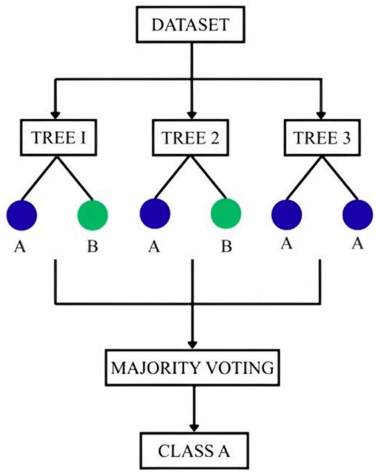



**Fig. 1.** Illustration of the RF Concept. Adapted from [29].

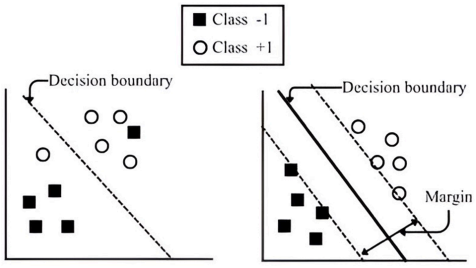



**Fig. 2.** Illustration of the SVM Concept. Adapted from [31].

fication is determined based on the majority vote (Fig. 1) [30].

The SVM model determines the optimal hyperplane position to separate the data classes (classes -1 and +1) in Fig. 2 [14], by observing the margin as the distance between the hyperplane and the nearest data point of each class [32]. The implementation of SVM uses three kernels, namely linear, RBF, and polynomial, to model and handle the complexity of hematological characters.

### 2.4 Evaluation of results

Models can be evaluated based on the values of accuracy, precision, recall, and F1-score [23].

| Actual \ Predicted | A | B | C |
|---|---|---|---|
| A | TP | FN | FN |
| B | FP | TN | TN |
| C | FP | TN | TN |

**Fig. 3.** Reading scheme of a multiclass confusion matrix (3x3). TP: True Positive, FN: False Negative, TN: True Negative, FP: False Positive. Adapted from [33].

Model performance was rated from a confusion matrix which displays the classification results information related to the t-value positive, false negative, false positive, and true negative (Fig. 3).

The values of accuracy, precision, recall, and F1-score are obtained from the values listed in the confusion matrix using the following formula [33]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (2.1)$$

$$Precision = \frac{TP}{TP + FP}, \quad (2.2)$$

$$Recall = \frac{TP}{TP + FN}, \quad (2.3)$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \quad (2.4)$$

The final step is to test the model using test data to obtain prediction results [34]. Data testing was used to measure the model's ability to produce predictions based on data that has never been used before. The test results showed differences in performance between models and the best-performing model was identified [35]. This study adds the stage of model testing using new data outside of training and testing data as an approach to assessing the generalization of the model so that it can show the performance of the model when applied in the real world.

## 3. Results and Discussion

### 3.1 Data characteristics

The data characteristics were examined as the initial step in modeling thalassemia screening using machine learning by assessing the distribution of hematological, categorical, and balance factors within the dataset. The data distribution across categories exhibits considerable imbalance; hence, a combination of SMOTE and undersampling was employed to rectify these issues. Dataset imbalances provide not only a technical challenge in modeling but also reflect a real clinical scenario by integrating several types of thalassemia and non-thalassemia groups.

The data characteristics are illustrated in Fig. 4(a), which presents the boxplot distribution of the selected hematological variables dataset. The hematological characteristics exhibited heterogeneous values that reflect the physiological conditions of both healthy individuals and those affected by thalassemia and other hemoglobinopathies.

The hemoglobin value distribution largely remained within the normal range; however, there were outliers below the normal range, indicating the pathophysiological condition of patients with reduced oxygen capacity. The erythrocyte distribution exhibited consistent variation, with only a few outliers at both the upper and lower extremes. Elevated erythrocyte counts reflect the body's adaptation to hypoxia, whereas reduced erythrocyte levels may indicate anemia. Genetic disorder induces hemolytic anemia, leading to increased erythropoietin levels that stimulate erythropoiesis, thereby augmenting erythrocyte production [36]. The hematocrit exhibited significant variation with several outliers at the lower threshold, suggesting anemia due to a reduction in erythrocyte count in the body.

The MCV and MCH exhibited a broad data distribution; the values are at or below the normal threshold, with the median being at the normal limit. All outliers exhibiting below-normal levels in both features suggested that certain individuals possessed hypochromic microcytic hematologic profiles [37]. The mean and range values of both parameters are relatively high because the dataset is dominated by samples from the normal group. Consequently, the distributions of MCV and MCH values are more concentrated toward the higher range compared to the other groups. MCHC remains within the normal range due to its relative stability compared to other parameters. Several outliers exhibit mild hypochromic conditions in individuals with thalassemia or other disorders.

### 3.2 Hematological features that affect thalassemia

The most relevant hematological features in differentiating thalassemia within the RF and SVM models comprise MCV, MCH, hemoglobin, hematocrit, and erythrocytes. These features were selected based on their clinical relevance, allowing the model to differentiate between thalassemia and non-thalassemia. These specific hematological parameters are physiologically associated with thalassemia. Thalassemia is a disorder related to hemoglobin in erythrocytes, making these hematologic parameters potential markers for its identification.

MCV is a measure that reflects the estimate of erythrocytes, while MCH is a measure that surveys the hemoglobin content of erythrocytes. Individuals with thalassemia are typically characterized by decreases in MCV and MCH, reflecting microcytic and hypochromic characteristics.

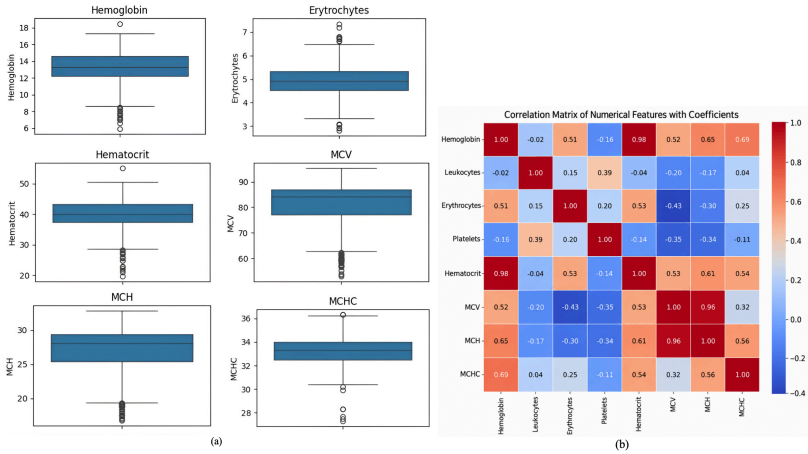



**Fig. 4.** (a) Boxplot showing the distribution, median, and presence of outliers in selected hematologic features, (b) Correlation matrix describing the relationship between hematological feature.

Decreased hemoglobin concentrations are associated with disruptions in globin chain synthesis. Hematocrit represents the ratio of erythrocytes to the total blood volume [39].

The hematocrit parameter is directly proportional with erythrocyte count; thus, variations in erythrocyte count will affect hematocrit values. Changes in erythrocyte count contribute to the identification of thalassemia. Erythrocytosis is regarded as a hematological feature of thalassemia patients, serving as a compensatory reaction to microcytosis and hypochromia [40]. In thalassemia patients, decreased bone marrow erythrocyte production, increased erythrocyte count, and precursor cell destruction ultimately result in anemia [41].

The selected features align with the pathophysiological characteristics of thalassemia, enhancing the precision of the machine learning model's performance. While metrics like MCV and MCH are critical indications for suspecting thalassemia, the inclusion of additional characteristics enhances the potential for unbiased models and augments the model's generalized ability to predict novel data. The incorporation of these variables aligns the model more closely with real-world diagnostic techniques [42]. Furthermore, this study incorporates the hybrid unsupervised-supervised learning approach for thalassemia screening in targeted regional populations. The unsupervised learning method was utilized to investigate the fundamental patterns and clustering of hematological data, while the supervised learning techniques i.e., RF and SVM were applied for classification.

### 3.3 Clustering using K-means and gaussian mixture model

The clustering results (Fig. 5) provide a preliminary overview of the hematological data separation patterns. Although this separation does not rely on category labels, it reveals a pattern in hematological data that support category differentiation and identify abnormalities or outliers within the established cluster.

Analysis of hematological parameter values in each K-Means cluster revealed distinct ranges of values across clusters (Table 3). Thalassemia major had the most data ($n$ = 9) in Cluster 0, which displays het-

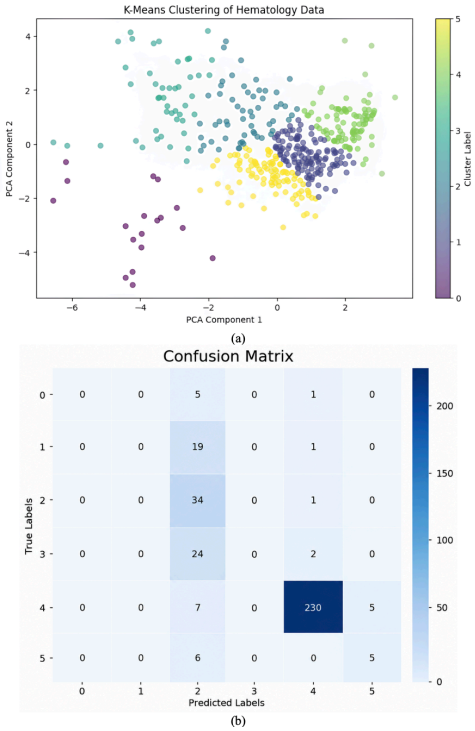



**Fig. 5.** A clustering visualization that divides the dataset into six clusters. (a) The results of K-means clustering are displayed in the form of scatter plots, (b) The results of GMM clustering are displayed through the confusion matrix.

**Table 2.** Category distribution per cluster.

| Categories | IDA | $\alpha$-thalassemia | $\beta$-thalassemia | HbE | Normal | Thalassemia Major |
|---|---|---|---|---|---|---|
| **Cluster** | | | | | | |
| 0 | 1 | 3 | 1 | 2 | 2 | 9 |
| 1 | 0 | 1 | 0 | 2 | 119 | 0 |
| 2 | 4 | 9 | 8 | 19 | 16 | 0 |
| 3 | 1 | 7 | 33 | 2 | 1 | 2 |
| 4 | 0 | 3 | 0 | 3 | 85 | 0 |
| 5 | 3 | 4 | 3 | 4 | 78 | 0 |

erogeneity in data across all groupings, but in very small numbers. The Hb, erythrocyte, hematocrit, and MCHC values in this cluster were the lowest among the clusters, whereas the MCV and MCH values are the second lowest when compared to all clusters. Consequently, this cluster is primarily characterized by thalassemia major. However, the upper limit values for multiple parameters fall within the normal range, allowing for the inclusion of other mild disorders in the group.

Clusters 1, 4, and 5 (Table 2) represent the normal data groups, indicated by their dominance and reflected in the mean hematological value being at the normal limit. The parameter ranges generally fell within normal limits, although some lower-bound values were already below the normal range. This may result in a slight reduction in the parameter values, causing data from different groups to overlap. In Cluster 5, the lower-bound values are lower than those in Clusters 1 and 4, resulting in greater variation within the cluster and increasing the likelihood of overlap with other abnormality groups. The normal category dominates these three clusters, which indicates the existence of sub-patterns or variations in the normal group. These variations are characterized by the differences in the value ranges of each parameter across the three clusters, although the ranges remain within the normal thresholds. This variability is likely due to the absence of data grouping by age or gender, as hematological parameters can differ by age groups or genders.

Cluster 2 exhibits significant data heterogeneity, with all category groups represented in varying properties. The normal and HbE groups dominate this cluster. The values range from subnormal to normal, allowing the inclusion of the $\beta$-thalassemia, $\alpha$-thalassemia, and IDA classifications (Table 3). A previous study indicated that hematological values in normal and HbE individuals were relatively similar, with HbE individuals ranging from asymptomatic to those requiring transfusions [43]. This mild HbE condition aligns with normal individuals when considering only hematological screening outcomes. Moreover, from the perspective of red blood cell morphology,

**Table 3.** The average value and range of hematological features of each cluster.

| Cluster | Hemoglobin | | | Leukocytes | | | Erythrocytes | | | Platelets | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | min | | | min | | | min | max | mean | min | max | mean |
| 0 | 5.9 | 9.7 | 7.89 | 1.21 | 16.82 | 6.57 | 2.81 | 4.64 | 3.55 | 23 | 495 | 238.61 |
| 1 | 12.8 | 15.1 | 13.85 | 4.50 | 13.90 | 7.74 | 4.16 | 5.55 | 4.78 | 175 | 459 | 294.45 |
| 2 | 10.6 | 14.9 | 12.91 | 4.76 | 18.06 | 9.43 | 4.31 | 7.35 | 5.29 | 185 | 576 | 368.98 |
| 3 | 7.7 | 13.4 | 10.92 | 5.23 | 14.51 | 8.86 | 4.08 | 6.81 | 5.59 | 271 | 645 | 402.5 |
| 4 | 14.5 | 18.5 | 15.82 | 4.24 | 19.89 | 7.65 | 4.67 | 6.71 | 5.42 | 191 | 608 | 289 |
| 5 | 10.4 | 13.3 | 12.32 | 3.85 | 13.42 | 6.98 | 3.65 | 5.08 | 4.43 | 158 | 431 | 292.98 |
| Cluster | Hematocrit | | | MCV | | | MCH | | | MCHC | | |
| | min | max | mean | min | max | mean | min | max | mean | min | max | mean |
| 0 | 19.7 | 32.1 | 25.72 | 56.8 | 87.6 | 72.84 | 17.6 | 27.8 | 22.4 | 27.3 | 33.4 | 30.75 |
| 1 | 37.5 | 46.0 | 41.36 | 75.1 | 94.7 | 86.48 | 25.6 | 31.7 | 28.96 | 31.5 | 36.2 | 33.49 |
| 2 | 31.4 | 44.7 | 38.99 | 59.6 | 87.9 | 74.22 | 19.3 | 29.4 | 24.58 | 31.2 | 34.8 | 33.12 |
| 3 | 24.9 | 40.3 | 33.95 | 53.0 | 78.2 | 61.11 | 16.8 | 25.0 | 19.63 | 28.3 | 34.9 | 32.14 |
| 4 | 40.8 | 55.0 | 46.01 | 76.8 | 92.5 | 85.07 | 25.9 | 32.8 | 29.26 | 32.5 | 36.3 | 34.4 |
| 5 | 29.4 | 40.5 | 37.6 | 72.5 | 95.3 | 84.97 | 23.4 | 32.6 | 27.85 | 30.6 | 35.7 | 32.79 |

HbE reveals similarities to individuals with thalassemia or mild thalassemia intermedia conditions.

Cluster 3 is dominated by the $\beta$-thalassemia group, while the other groups are represented in smaller proportions. This is reflected in the parameter ranges, which are generally lower than those observed in the other clusters, except for Cluster 1. The slightly lower values may allow the IDA, $\alpha$-thalassemia, HbE, and normal samples within this cluster. This cluster exhibited a characteristic profile of $\beta$-thalassemia, as indicated by the decreased range of hemoglobin and hematocrit values, as well as the lowest MCV and MCH values observed across all clusters. Conversely, erythrocyte levels were significantly increased.

Clusters 0 and 2 indicate that the data patterns remain relatively ambiguous, making it difficult to clearly distinguish between the groups. The hematological value ranges within certain clusters included both normal values and values below the normal thresholds. Consequently, the clusters illustrate a blend of categories, indicating the limitations of hematological data patterns to differentiate categories clearly.

Clustering using GMM group data based on probability distributions was performed. In contrast to K-means, which is commonly visualized in a scatter plot, GMM clustering in this study is presented using a confusion matrix to compare cluster assignment with the actual hematological labels. The diagonal cells of the confusion matrix indicate the number of correct classifications. Cluster 4 shows the highest number of correctly classified instances, predominantly composed of normal data (230 instances) and followed by cluster 2, mainly consisting of $\beta$-thalassemia (34 instances). Although the visualization approaches differ, both clustering methods exhibit that the normal and $\beta$-thalassemia groups possess sufficiently distinct data patterns compared to other groups. The accuracy of the GMM model at 79.11%, was determined by comparing the projected result clusters with the actual labels.

Meanwhile, the $\beta$-thalassemia category appears to overlap with $\alpha$-thalassemia and HbE. This is likely due to a mild decrease in $\beta$-thalassemia-related hematological parameters, which may result in the inclusion of $\alpha$-thalassemia and HbE data.

In another study, it was shown that the hematological profiles of the $\beta$-thalassemia trait and homozygous $\alpha$-thalassemia were highly similar due to their microcytic characteristics, with no statistically significant differences in red cell indices between the $\beta$-thalassemia trait and homozygous $\alpha$-thalassemia [44]. A previous study on the difference in hematological profiles of HbE and $\beta$-thalassemia found that the hematological parameters of $\beta$-thalassemia are generally lower than HbE, although the HbE values also tend to fall below normal thresholds [45].

Clustering is a method used to identify inherent patterns of hematological data distribution within each clinical category. This data analysis reveals the limitations of overlapping hematological data patterns that underpin the evaluation of prediction outcomes in the model test.

### 3.4 Evaluation results of random forest and support vector machine models

The evaluation of model performance evaluation was conducted to compare the RF and SVM models with linear, RBF, and polynomial kernels. Comparisons were conducted by analyzing the accuracy, precision, recall, and F1-score values, as well as the confusion matrix data to assess the model's ability to distinguish hematological categories. The hematological features utilized demonstrate a strong diagnostic ability in distinguishing cases of thalassemia from normal individuals, as reflected in the model performance results displayed in Table 4.

Previous studies have stated that an accuracy above 70% indicates that the algorithm's performance is sufficiently good [46]. According to this threshold, all models developed this study demonstrated satisfactory performance. The RF model achieved the highest accuracy at 91%, whereas with the SVM model, the linear kernel demonstrated the highest accuracy value of 86%, which was only slightly higher than the RBF kernel (85%).

The RF model exhibited strong performance across all model evaluation metrics, indicating that the RF model exhibits considerable resilience to the complexity of hematological features. This resilience is achieved by integrating multiple trees, each trained on distinct random subsets of the dataset. Consequently, misclassifications were mitigated, as the aggregation of multiple trees offered diverse decisions, thereby enhancing predictive accuracy [47]. SVM demonstrated stable and reliable performance (Table 4) on small datasets characterized by complex patterns, utilizing several kernel functions to construct a flexible decision boundary for data mapping. These findings continue to support the SVM for achieving competitive accuracy under specific situations [48].

**Table 4.** Evaluation results of the RF and SVM models.

| Models | RF | SVM | | |
|---|---|---|---|---|
| Kernel | - | Linear | RBF | Polynomial |
| Accuracy | 91% | 86% | 85% | 76% |
| Precision | 89% | 89% | 89% | 87% |
| Recall | 86% | 86% | 85% | 76% |
| F1-score | 87% | 87% | 86% | 80% |

SVM-linear is appropriate for use on data with linear inter-feature relationships, whereas SVM-RBF is better suited for handling non-linear and more complex data patterns, and SVM-polynomial is best utilized to model feature combinations or interactions [24]. The difference in precision and recall metrics across the RF and SVM models indicates an imbalance in class predicting. Nonetheless, the comparable F1-score values on both metrics indicate that the balance is still maintained.

Previous studies have illustrated the differenced in effectiveness between these two models, each yielding distinct outcomes. These findings indicate that both the dataset properties and the applied optimizations affect the model's performance. In another study using models to predict $\beta$-thalassemia, the RF model attained an accuracy of 91%, while the SVM model achieved an accuracy of 90% [49]. Misclassified instances as false positives indicate a pattern within a proximate or intersecting range. Employing SVM with a linear kernel and RBF showed consistent efficacy in classifying patients with $\alpha^0$-thalassemia and $\alpha^+$-thalassemia based on CBC data, achieving an accuracy of 90.72% [42]. Using SVM-polynomial achieved 99% accuracy in the classification of $\alpha$-thalassemia and $\beta$-thalassemia in both minor and major types [50]. In the investigation of the efficacy of RF and SVM models utilizing both original data and data enhanced by SMOTE intervention, performance improved after SMOTE was applied; the accuracy of the RF model increased from 92% to 95%, while the accuracy of the SVM increased from 91% to 92% [51]. The performance variability of these two models indicates that the algorithm's performance was influenced by the characteristics of the data, the optimization strategies applied, and the differences in the CBC parameters used. Nevertheless, the results obtained have shown that the machine learning approach in these two models was able to capture relevant patterns in hematological data and has the potential to be an effective screening tool.

The RF model demonstrated superior performance, indicating that hematological data patterns can be effectively distinguished using the decision tree methodology. Although, the SVM model showed lower performance overall, the SVM kernel suggests that the hematological data is sufficiently complex to be effectively separated in linear and nonlinear spaces utilizing the comparable performance of the linear and RBF kernels.

### 3.5 Test results of random forest and support vector machine models using new data

The performance evaluation of the RF and SVM models utilizing additional test data revealed differences in prediction results (Tables 5-6). These variations are reflected by the difference in the number of correct predictions obtained in each model. The SVM model had the highest number of correct predictions with a linear kernel of 40 data points, while the RF model has a total of 38 data points.

**Table 5.** Prediction results using the RF model.

| Categories | Prediction Results |
|---|---|
| IDA | 1 IDA |
| | 1 $\alpha$-thalassemia |
| $\alpha$-thalassemia | 5 $\alpha$-thalassemia |
| | 1 Normal |
| | 2 HbE |
| | 2 $\beta$-thalassemia |
| $\beta$-thalassemia | 7 $\beta$-thalassemia |
| | 2 $\alpha$-thalassemia |
| | 1 Thalassemia Major |
| HbE | 7 HbE |
| | 2 $\alpha$-thalassemia |
| | 1 $\beta$-thalassemia |
| Normal | 10 Normal |
| Thalassemia major | 8 Thalassemia Major |
| | 1 HbE |
| | 1 $\alpha$-thalassemia |
| **Total Correct Prediction** | **38** |

The test results of all models showed that only the normal hematological category was predicted to be 100% correct. Hematological profiles in normal individuals (hemoglobin, MCV, MCH, erythrocytes, and hematocrit) generally have clearer and more stable patterns, making them easier to distinguish across models than other hematological categories.

**Table 6.** Prediction results using the SVM model.

| Categories | Prediction Results | | |
|---|---|---|---|
| | SVM-linear | SVM-RBF | SVM-polynomial |
| IDA | 2 IDA | 2 IDA | 2 IDA |
| $\alpha$-thalassemia | 2 $\alpha$-thalassemia<br>3 IDA<br>3 HbE<br>2 $\beta$-thalassemia | 4 $\alpha$-thalassemia<br>1 IDA<br>3 HbE<br>2 $\beta$-thalassemia | 8 IDA<br>2 $\beta$-thalassemia |
| $\beta$-thalassemia | 9 $\beta$-thalassemia<br>1 $\alpha$-thalassemia | 7 $\beta$-thalassemia<br>1 HbE<br>1 $\alpha$-thalassemia<br>1 Thalassemia Major | 9 $\beta$-thalassemia<br>1 IDA |
| HbE | 8 HbE<br>2 $\alpha$-thalassemia | 6 HbE<br>3 $\alpha$-thalassemia<br>1 IDA | 4 HbE<br>6 IDA |
| Normal | 10 Normal | 10 Normal | 10 Normal |
| Thalassemia major | 9 Thalassemia Major<br>1 HbE | 3 Thalassemia Major<br>7 HbE | 5 Thalassemia Major<br>5 HbE |
| **Total Correct Prediction** | **40** | **32** | **30** |

Classification of $\beta$- thalassemia is recognized with various performance. The SVM-linear and SVM-polynomial showed good classification performance (9 correctly classified instances), whereas RF and SVM-RBF demonstrated moderate performance (7 correctly classified instances). Misclassification led to the categories of $\alpha$-thalassemia, thalassemia major, HbE, and IDA. The similarity of hematological profiles between $\beta$-thalassemia and HbE have the potential to cause overlap, as they involve abnormalities in the $\beta$-globin gene, resulting in relatively adjacent patterns of hematologic parameters. HbE can give rise to a $\beta$-thalassemia-like picture due to a substitution mutation in codon 26 of the $\beta$-globin gene that changes the sequence GAG (glutamate) to AAG (Lysine). This condition causes cryptic splice site activation thereby decreasing the production of mRNA $\beta$E and reducing the synthesis of the $\beta$E chain of globin [52]. In addition, thalassemia major in this study includes heterozygous $\beta$-thalassemia and HbE so it also has the potential to show an overlapping hematologic profile because it involves abnormalities in the same gene. When compared to IDA, hemoglobin values are in a close range, only in $\beta$-thalassemia is it higher. Meanwhile, the values of MCV and MCH in $\beta$-thalassemia are lower than those of IDA.

$\alpha$-thalassemia exhibited limited classification performance across the models. RF showed moderate performance (5 correctly classified instances), the linear and RBF SVM showed lower performance (2-4 correctly classified instances). Prediction errors led to normal, HbE, $\beta$-thalassemia, and IDA classifications. The data suggests that the hematologic pattern in $\alpha$-thalassemia has less prominent characteristics, making it more difficult to distinguish by the model. Hematological parameters in $\alpha$-thalassemia typically exhibit a mild decrease, which may resemble the conditions of IDA, other hemoglobinopathies, or even normal levels. Mild deterioration in conditions such as $\beta$-thalassemia and HbE can lead the model to misidentify cases as $\alpha$-thalassemia.

Furthermore, the less prominent characteristics of $\alpha$-thalassemia might be caused by the lack of subgroup information in the archived dataset. The $\alpha$-thalassemia cases were not further categorized into $\alpha$0 and $\alpha$+ thalassemia subgroups. This may have increased the variability of the $\alpha$-thalassemia group and decreased the specificity of its hematological pattern because samples with potentially distinct hematological symptoms were combined into one category. This limitation may have contributed to the lower classification performance observed for $\alpha$-thalassemia. The findings of a separate study suggest that the

MCV and MCH values of $\alpha$-thalassemia trait carriers are in the normal range to slightly decreased, though confirmation of the diagnosis often requires DNA analysis [53]. Furthermore, another study showed that silent $\alpha$-thalassemia carriers had a normal hematological profile, showed no significant differences, and usually had HPLC test results that showed normal $\alpha$-globin levels [54]. This condition indicates that silent $\alpha$-thalassemia carriers are difficult to identify using hematologic parameters.

The HbE category was predicted with the highest performance by the SVM-linear model (8 correctly classified instances, followed by the RF (7 correctly classified instances) and then the SVM-RBF (5 correctly classified instances), with the lowest performance seen in the SVM-polynomial (3 correctly classified instances) model. Misclassifications indicate categories apart from HbE, including $\alpha$-thalassemia, $\beta$-thalassemia, and IDA. The range of hemoglobin, MCV, and MCH parameter values in the HbE category is relatively close to the thalassemia group. However, the hematological mean value of HbE was the highest compared to the thalassemia group, while the $\beta$-thalassemia category showed the lowest average value. In IDA, the MCV and MCH values were slightly higher, while Hb was lower than the thalassemia group. IDA had a lower MCH but a higher MCV when compared to HbE. This overlap of hematological parameters makes it difficult for the model to distinguish categories.

The IDA data were correctly predicted by all SVM kernels, whereas the RF correctly classified only one instance. Misclassifications led to $\alpha$-thalassemia classification due to overlapping hematological values, where IDA generally exhibits lower hemoglobin values but higher MCV and MCH values compared to $\alpha$-thalassemia. The thalassemia major data was correctly classified most by SVM-linear ($n$ = 9), followed by RF (n = 8), and moderately well by SVM-polynomial ($n$ = 5), while the lowest performance was seen in SVM-RBF ($n$ = 3). Misclassification occurred mostly in the HbE category, with one instance misclassified as $\alpha$-thalassemia. The inaccuracy of prediction for anemia and thalassemia groups can be explained by the overlapping of hematological characteristics, since patients with low MCV and MCH values generally present with microcytic and hypochromic anemia. Furthermore, the misclassification of HbE might be related to the presence of heterozygous $\beta$-thalassemia and HbE in some thalassemia major samples. Although certain hematological parameters in thalassemia major (hemoglobin, MCV, MCH) are sufficient to show a clear enough pattern—the lowest mean and range—to be separated from other categories.

Model evaluation using independent test data, separate from training and testing data, provides an external validation stage to assess the model's generalization capabilities and potential for implementation in the real world.

This preliminary study applied a combination of unsupervised and supervised learning to the dataset. Unsupervised learning was used to explore patterns in hematological data among the thalassemia group, other hemoglobinopathies, and normal individuals. The hematological data patterns revealed a tendency for clustering among certain groups, particularly between the normal and $\beta$-thalassemia groups, although some overlap still remained. Overlap was also found between the HbE and normal groups, as well as among other

groups. Therefore, consistent separation patterns based on hematological parameters were not clearly observed. Supervised learning contributed to the development of RF and SVM prediction models using linear, RBF, and polynomial kernels to classify new test data that were not included in the training set as an evaluation of the models' generalization ability to previously unseen data.

### 3.6 Clinical diagnostic limitations

The results of this preliminary study indicate that hematological profiles have limitations in distinguishing between patients with thalassemia and those with other hemoglobinopathies. The similar characteristics of microcytic anemia in these hematological profiles make it difficult to differentiate between the thalassemia, HbE, and IDA groups. A CBC is very useful as an initial screening tool but cannot yet serve as a definitive diagnostic tool, thus requiring further definitive diagnostic testing. The overlap in hematological profiles observed in this analysis aligns with the challenges encountered in real-world clinical settings during the diagnostic process. CBC results showing a decrease in Hb levels still require additional testing to distinguish between conditions such as thalassemia and IDA. Therefore, screening results based on a CBC must be followed up with more specific tests according to clinical indications to differentiate thalassemia IDA.

This overlap may be related to variations in hematological characteristics in mild forms of thalassemia, which may show relatively minor changes in the hematological profile, resulting in similar values. For example, mild $\beta$-thalassemia may exhibit hematological characteristics similar to IDA [40, 55]. Similarly, a microcytic profile can be found in $\beta$-thalassemia, $\alpha$-thalassemia, and IDA.

Limitations in the interpretation of $\alpha$-thalassemia groups may also be related to the classification of groups within the dataset. Variations in $\alpha$-thalassemia genotypes result in variations in hematological characteristics; however, some conditions remain difficult to distinguish based solely on hematological profiles [56]. In this study, the categories from the available sources did not include information regarding the $\alpha 0$ and $\alpha+$ genotype subgroups; therefore, these categories were combined into a single dataset class.

Age variation in the dataset may have also contributed to the wide range of hematological profile values. The sample included individuals ranging from children to adults to the elderly, given that normal CBC parameter values differ across age groups. This factor must be considered when interpreting classification results if the data distribution shows a wide range.

## 4. Conclusion

The evaluation results show that RF achieved the highest accuracy (91%). However, when testing data not used in the training process, SVM-linear demonstrated the highest performance, with 40 correctly predicted instances, although this was only two instances higher than RF. These results indicate that RF exhibits superior classification performance based on accuracy, whereas SVM-linear demonstrates slightly better generalization ability on the test data.

This study still has limitations regarding the sample size and the availability of supporting information in the dataset. Future research is recommended to use a larger dataset and incorporate additional supporting parameters to help clarify the patterns between groups, thereby potentially improving model performance and

generalization.

## Acknowledgements

The authors would like to acknowledge the interdisciplinary collaboration between biological science and computational intelligence, which facilitated this research. This research received no specific grant from any funding agency.

## References

[1] Paloma IDANC. Thalassemia: a literature review. Biocity Journal of Pharmacy Bioscience and Clinical Community. 2023;1(2):89-100.

[2] Tuo Y, Li Y, Li Y, Ma J, Yang X, Wu S, et al. Global, regional, and national burden of thalassemia, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. EClinicalMedicine. 2024;72.

[3] Mendri NK, Prayoga SA. Nursing care for sick children and high-risk infants. Yogyakarta: Pustaka Baru Press; 2017.

[4] Hertini R, Kurniati E, Suriany S. Self-description with the quality of life of thalassemia patients in the children's ward of the Banjar City Hospital. Jurnal Keperawatan Galuh. 2024;6(1):87-94.

[5] Kattamis A, Forni GL, Aydinok Y, Viprakasit V. Changing patterns in the epidemiology of $\beta$-thalassemia. European Journal of Haematology. 2020;105(6):692-703.

[6] Rujito L, Maritska Z, Sofro AS. $\beta$-Thalassemia mutation flow in Indonesia: a migration perspective. Thalassemia Reports. 2023;13(4):253-261.

[7] Ministry of Health of the Republic of Indonesia. Decree of the Minister of Health of the Republic of Indonesia Number HK.01.07/MENKES/1/2018 concerning national guidelines for thalassemia management medical services. Jakarta: Ministry of Health of the Republic of Indonesia; 2018.

[8] Rujito L. Buku referensi talasemia: genetik dasar dan pengelolaan terkini. Purwokerto: Universitas Jenderal Soedirman; 2019.

[9] Directorate General of Disease Prevention and Control, Directorate of Prevention and Control of Non-Communicable Diseases, Ministry of Health of the Republic of Indonesia. Pocket book of educational guidance for thalassaemia counseling and laboratory examination. Jakarta: Ministry of Health of the Republic of Indonesia; 2021.

[10] Husna N, Sanka I, Al Arif A, Putri C, Leonard E, Handayani NN. Prevalence and distribution of thalassemia trait screening. J Med Sci. 2017;49(3):106-113.

[11] Tjahyanti LPAS, Saputra PS, Gitakarma MS. The role of artificial intelligence (AI) to support learning during the COVID-19 pandemic. Context. 2022;1(1).

[12] Zubaidah, Panggabean T, Alvito P, Akbar Z, Nadia CM. Artificial intelligence systems in bioinformatics: future trends and potential. Jurnal Sistem Informasi dan Ilmu Komputer. 2025;3(1):24-31.

[13] Putri M, Erlin. Stroke prediction using machine learning with a random forest algorithm. Infomedia Journal: Informatics, Multimedia, and Network Engineering. 2024;9(1):16-21.

[14] Kusuma PD. Machine learning theory, programs, and case studies. Yogyakarta: Deepublish; 2020.

[15] Heryadi Y, Wahyono T. Machine learning: concept and implementation. Yogyakarta: Gava Media; 2020.

[16] Gao CX, Dwyer D, Zhu Y, Smith CL, Du L, Filia KM, et al. An overview of clustering methods with guidelines for applica-

tion in mental health research. Psychiatry Research. 2023;327:115265.

[17] Faouzi J, Colliot O. Classic machine learning methods. In: Colliot O, editor. Machine learning for brain disorders. Cham: Humana; 2023.

[18] Ferih K, Elsayed B, Elshoeibi AM, Elsabagh AA, Elhadary M, Soliman A, et al. Applications of artificial intelligence in thalassemia: a comprehensive review. Diagnostics. 2023;13(9):1551.

[19] Patricia S, Marpaung CP, Wijaya LR, Paramartha MA, Atmadja WD, Ningsih RY. Implementation of exploratory data analysis (EDA) to analyze various risk factors for heart disease in the United States. Journal of Student Development Information System (JoSDIS). 2023;3(2):108-124.

[20] Devanath A, Akter S, Karmaker P, Sattar A. Thalassemia prediction using machine learning approaches. In: Proceedings of the 6th International Conference on Computing Methodologies and Communication (ICCMC); 2022. p. 1166-1174.

[21] Mambang M, Hidayat A, Marleny FD, Wahyudi J. Exploratory data analysis to evaluate keyword search for learning videos on YouTube with a machine learning approach. Journal of Informatics and Computer Technology (JITEK). 2022;2(2):181-189.

[22] AlShammari AF. Implementation of feature selection using correlation matrix in Python. International Journal of Computer Applications. 2024;97:8887.

[23] Siregar AP, Purba DP, Pasaribu JP, Bakara KR. Implementation of random forest algorithm in the classification of stroke disease diagnosis. Journal of Engineering Sciences Research. 2023;2(4):155-164.

[24] Saifuddin, Azhari L, Widarti E, Wartono. Evaluation of the performance of linear kernels, RBF, and polynomial in a support vector machine model for hypertension risk prediction. FIFO Scientific Journal. 2025;17(2):192-201.

[25] Sabilla WI, Vista CB. Implementation of SMOTE and undersampling on imbalanced datasets for predicting company bankruptcy. Journal of Applied Computing. 2021;7(2):329-339.

[26] Nasution AU, Ambarita EP, Nurhidayanti N, Fadlia HE. Application of the Gaussian mixture models (GMM) algorithm to identify error patterns in compilation. JOMLAI: Journal of Machine Learning and Artificial Intelligence. 2025;4(2):88-98.

[27] Hutagalung MI, Sriani S. Clustering of ENT disease data using K-means clustering algorithm. MALCOM: Indonesian Journal of Machine Learning and Computer Science. 2024;4(4):1568-1577.

[28] Putra INAM, Pramartha C. Hyperparameter optimization of the support vector machine algorithm in the classification of $\beta$-thalassemia disease. National Journal of Information Technology and Its Applications. 2025;3(2):283-294.

[29] Khan MY, Qayoom A, Nizami MS, Siddiqui MS, Wasi S, Raazi SMKUR. Automated prediction of good dictionary examples (GDEX): a comprehensive experiment with distant supervision, machine learning, and word embedding-based deep learning techniques. Complexity. 2021;2021:2553199.

[30] Rahmaddeni, Wulandari D, Aziz S, Adrianto S, Pratiwi F. Theory and implementation of machine learning using Python. Payakumbuh: PT Serasi Media Teknologi; 2025.

[31] Buntoro GA. Sentiment analysis of DKI Jakarta governor candidates 2017 on Twitter. INTEGER: Journal of Information Technology. 2017;2(1).

[32] Putrawansyah F, Susanti T. Application of the support vector machine method to the classification of guava types. JIKO (Journal of Informatics and Computers). 2024;8(1):193-204.

[33] Wabang K, Nurhayati OD. Application of the naive Bayes classifier algorithm to classify community complaints. RESTI Journal (Systems Engineering and Information Technology). 2022;6(5):872-876.

[34] Al Hafiz DM, Amaly K, Jonathan J, Rachmatullah MT. The heart disease prediction system uses the naive Bayes method. Sriwijaya Electrical Engineering Journal. 2021;2(2).

[35] Alfianti MLT, Supriyanto R. Comparison of the performance of random forest, AdaBoost, and XGBoost algorithms in predicting the risk of osteoporosis. Journal of Computer Science and Agri-Informatics. 2024;11(2):172-184.

[36] Arna YD, Bobaya J, Wulandari D, Dwi NHR, Wahab I, Andriani Y, et al. Hematology. Cilacap: PT Media Pustaka Indo; 2025.

[37] Lestari P, Astuti D, Maharani EA. Anemia screening for medical laboratory technology students of the Ministry of Health Polytechnic of the Ministry of Health Jakarta III. Anakes: Scientific Journal of Health Analysts. 2024;10(1):69-76.

[38] Azwan A, Rahmad K, Pizaini P, Fitri I. Application of K-means clustering and correlation matrix algorithm to analyze the risk of dengue fever spread in the city of Pekanbaru. Independent Informatics Journal of Pasuruan. 2021;6(3):1-6.

[39] Ministry of Health of the Republic of Indonesia. Data interpretation guidelines clinic. Jakarta: Ministry of Health of the Republic of Indonesia; 2011.

[40] Urrechaga E, Borque L, Escanero JF. Erythrocyte and reticulocyte parameters in iron deficiency and thalassemia. Journal of Clinical Laboratory Analysis. 2011;25(3):223-228.

[41] Chaudhry AF, Malik Z, Shegos CJ. Rare co-inherited alpha-thalassemia minor and beta-thalassemia minor with heterozygous H63D mutation mistaken as iron deficiency anemia: a case report. AME Case Reports. 2022;6:4.

[42] Mohammadi E, Rastegar M, Jamshidnezhad A, Azizi A. Machine learning improves detection of alpha thalassemia carriers compared to clinical features. Scientific Reports. 2025;15(1):36717.

[43] Vichinsky E. Hemoglobin E syndromes. ASH Education Program Book. 2007;2007(1):79-83.

[44] Mehdi SR, Al Dahmash BA. A comparative study of hematological parameters of $\alpha$- and $\beta$-thalassemias in a high prevalence zone: Saudi Arabia. Indian Journal of Human Genetics. 2011;17(3):207.

[45] Nandi A, Talukdar M, Bhattacharya S, Sen S, Biswas S, Roy K. Red blood cell indices in different hemoglobinopathies: a cross-sectional study in Eastern India. Indian Journal of Pathology and Microbiology. 2024;67(1):115-120.

[46] Syahri A, Fariha U, Afandi R, Nurliyana I. Comparison of logistic regression, random forest and AdaBoost algorithms for diabetes mellitus classification. IJATIS: Indonesian Journal of Applied Technology and Innovation Science. 2024;1(1):41-46.

[47] Inonu OY, Magda K, Amarudin A. Performance analysis of random forest algorithms with machine learning model on diabetes dataset. Expert J Manaj Sist Inf dan Teknologi. 2025;15(1):1.

[48] Zuhria R, Habibi AR, Ullah S, Aslam W, Mehmood A, Choi GS, et al. Comparative analysis of random forest and SVM performance in asthma prediction. Sinkron:

Jurnal dan Penelitian Teknik Informatika. 2025;9(1):347-356.

[49] Sadiq S, Khalid MU, Ullah S, Aslam W, Mehmood A, Choi GS, et al. Classification of $\beta$-thalassemia carriers from red blood cell indices using ensemble classifier. IEEE Access. 2021;9:45528-45538.

[50] Nasir MU, Zubair M, Naseem MT, Shahzad T, Saeed A, Adnan KM, et al. Multiclass classification of thalassemia types using complete blood count and HPLC data with machine learning. Scientific Reports. 2025;15(1):26379.

[51] Rustam F, Ashraf I, Jabbar S, Tutusaus K, Mazas C, Barrera AEP, et al. Prediction of $\beta$-thalassemia carriers using complete blood count features. Scientific Reports. 2022;12(1):19999.

[52] Hirsch RE, Sibmooh N, Fucharoen S, Friedman JM. HbE/$\beta$-thalassemia and oxidative stress: the key to pathophysiological mechanisms and novel therapeutics. Antioxidants & Redox Signaling. 2017;26(14):794-813.

[53] Maharani EA, Astuti DA. Calculation of erythrocyte formula index in thalassemia minor screening test. Journal of Health Science and Technology. 2014;2(1):53-59.

[54] Uçucu S. Molecular spectrum and chromatographic-haematological characterisation of alpha thalassemia minor (-$\alpha$/-$\alpha$), and alpha silent carriers (-$\alpha$/-$\alpha\alpha$). medRxiv. 2023.

[55] Giordano PC. Strategies for basic laboratory diagnostics of the hemoglobinopathies in multi-ethnic societies: interpretation of results and pitfalls. International Journal of Laboratory Hematology. 2013;35(5):465-479.

[56] Akhavan-Niaki H, Kamangari RY, Banihashemi A, Oskooei VK, Azizi M, Tamaddoni A, et al. Hematologic features of alpha thalassemia carriers. International Journal of Molecular and Cellular Medicine. 2012;1(3):162.

## Appendix

**Table 7.** Hyperparameter settings used for each machine learning model in this study.

| Model | Hyperparameter | Value |
|---|---|---|
| **RF** | n_estimators | 100 |
| | criterion | Gini |
| | max_depth | None |
| | min_samples_split | 2 |
| | min_samples_leaf | 1 |
| | max_features | sqrt |
| | bootstrap | True |
| | random_state | 42 |
| **SVM-linear** | C | 1.0 |
| | kernel | linear |
| | random_state | 42 |
| **SVM-RBF** | C | 1.0 |
| | kernel | rbf |
| | gamma | scale |
| | random_state | 42 |
| **SVM-polynomial** | C | 1.0 |
| | kernel | poly |
| | degree | 3 |
| | gamma | scale |
| | random_state | 42 |