APPLICATION OF MACHINE LEARNING IN SCREENING FOR Β-THALASSEMIA USING COMPLETE BLOOD COUNT (CBC) PARAMETER

Ta Van Thao✉, Tran Hai Yen, Dang Thi Thuy Hong

Hanoi Medical University

This study is among the first in Vietnam to apply machine learning (ML) for β-thalassemia screening based solely on complete blood count (CBC) parameters. A regionally imbalanced dataset of 515 CBC samples was collected from students in Lai Châu province at Chemedic Laboratory (Hanoi) between October and December 2023. A validation set of 111 samples, including 55 β-thalassemia cases confirmed by high-performance liquid chromatography (HPLC), was analyzed using an XP-100 hematology analyzer. Supervised ML models-Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR)-were developed with Python libraries (scikit-learn, TensorFlow), incorporating Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), Principal Component Analysis (PCA), and Singular Value Decomposition (SVD) for data balancing and feature extraction. The SMOTE-PCA/SVD combinations achieved high accuracy (0.95 for DT and RF; 0.93 for LR), with ROC AUC 0.94–0.96 and F1-score ≈ 0.95. Using ADASYN with PCA/SVD improved DT accuracy to 0.97 but reduced RF to 0.85. Optimal performance occurred with a 500-sample training set, 60:40 class ratio, and test sizes of 0.05 – 0.2. These findings demonstrate that ML, particularly DT and RF, can serve as cost-effective, non-invasive screening tools for β-thalassemia in resource-limited regions of Vietnam, although further validation with larger and genetically confirmed datasets is warranted.

Keywords: β-thalassemia, machine learning, complete blood count, screening.

I. INTRODUCTION

Thalassemia is an inherited hemolytic anemia caused by genetic mutations that reduce or eliminate synthesis of α- or β-globin chains, potentially leading to severe anemia, organ damage, or early mortality.[1,2] Classified as α- or β-thalassemia depending on the affected globin chain, it represents a major global public health challenge due to its high prevalence and associated morbidity.[3,4] β-Thalassemia is particularly common in Southeast Asia, including Vietnam, where the carrier rate is estimated at approximately 7.8%.[1] According to the Thalassemia International Federation (2022), around 7% of the global population are carriers, while the World Health Organization (2008) reports that hemoglobinopathies affect 71% of countries, resulting in 60,000 – 70,000 births annually with severe β-thalassemia and contributing to 3.4% of under-5 mortality.[1,2]

Managing β-thalassemia imposes a substantial economic and healthcare burden due to lifelong blood transfusions and iron chelation therapy.[1] Early detection through cost-effective and accessible screening methods such as CBC analysis can facilitate timely interventions, reducing healthcare costs and improving outcomes.[5] Although confirmatory tests such as high-performance liquid chromatography (HPLC), hemoglobin electrophoresis, and molecular analysis remain

Corresponding author: Ta Van Thao
Hanoi Medical University
Email: tavanthao@hmu.edu.vn
Received: 15/09/2025
Accepted: 15/10/2025

the diagnostic gold standards, their high cost and technical complexity limit widespread use in low-resource settings.[6]

Recent advances in machine learning (ML) have leveraged CBC data to classify thalassemia and distinguish it from iron-deficiency anemia using supervised algorithms with data-balancing and dimensionality-reduction techniques such as SMOTE, ADASYN, PCA, and SVD.[7,8] While AI-based decision support systems for prenatal thalassemia screening have been explored in Vietnam, these approaches incorporated both hematological and biochemical variables. To our knowledge, no prior study in Vietnam has focused exclusively on applying ML to CBC-derived erythrocyte indices for β-thalassemia screening in a general population.[9] This study therefore aims to evaluate the performance of supervised ML models for β-thalassemia detection and to assess the effects of different data-balancing and feature-extraction methods (SMOTE, ADASYN, PCA, and SVD) using a small, regionally imbalanced dataset from a high-prevalence area.[10,11]

II. MATERIALS AND METHODS

1. Subjects

The study included a training dataset of 515 peripheral blood samples collected from individuals born in 2009 in Lai Chau province, Vietnam, selected from an initial cohort of 2,813 samples based on normal plasma iron and ferritin levels, conducted between October 2023 and December 2023 at Chemedic Laboratory, Hanoi. A validation dataset comprised 111 samples, with 55 confirmed β-thalassemia cases and 56 non-affected controls, diagnosed via high-performance liquid chromatography (HPLC).[2] The demographic profile included 274 males (53%) and 241 females (47%).

Inclusion Criteria: Samples collected in ethylenediaminetetraacetic acid (EDTA)-coated tubes with a minimum volume of 2 mL, free of hemolysis, and processed within 24 hours at 4°C.

Exclusion Criteria: Samples compromised during transportation, storage, or preservation; samples from individuals with iron-deficiency anemia (defined by MCV < 80fL and ferritin < 15 ng/mL).

2. Methods

This cross-sectional study employed purposive sampling based on the defined inclusion and exclusion criteria. The dataset consisted of 252 individuals diagnosed with β-thalassemia and 263 healthy controls, with disease status initially screened by complete blood count (CBC) and confirmed by HPLC. The dataset included 12 features: 9 hematological parameters (red blood cell count [RBC], hematocrit [HCT], hemoglobin [HGB], mean corpuscular volume [MCV], mean corpuscular hemoglobin [MCH], mean corpuscular hemoglobin concentration [MCHC], red cell distribution width [RDW], platelet count [PLT], white blood cell count [WBC]), 2 demographic variables (age and gender), and 1 binary classification target (β-thalassemia status), as detailed in Table 1.

Table 1. Features and Their Significance

Feature	Significance	Data Type
Age	Age of study participant	Numeric
Sex	Gender of diagnosed patient	Categorical
RBC	Red blood cell count	Numeric

Feature	Significance	Data Type
HCT	Hematocrit (packed cell volume)	Numeric
HGB	Hemoglobin concentration	Numeric
MCV	Mean corpuscular volume	Numeric
MCH	Mean corpuscular hemoglobin	Numeric
MCHC	Mean corpuscular hemoglobin concentration	Numeric
RDW	Red cell distribution width	Numeric
PLT	Platelet count	Numeric
WBC	White blood cell count	Numeric
Thalassemia	Diagnosis of β-thalassemia (yes/no)	Binary

Laboratory Procedures

Blood samples underwent the following sequential analyses:

Complete Blood Count (CBC): Performed using an XP-100 analyzer (Sysmex Corporation) per the Ministry of Health's guidelines for Hematology-Transfusion-Medicine-Immunology-Genetics-Molecular Biology Procedures. Venous blood (2 mL) was collected in EDTA-coated tubes, mixed thoroughly, and analyzed within 24 hours at 4°C. Reference ranges were: RBC (males: 4.2 – 5.4 T/L, females: 4.0 – 4.9 T/L), HCT (males: 0.4 – 0.47 L/L, females: 0.37 – 0.42 L/L), HGB (≥ 120 g/L), MCV (80 – 100fL), MCH (28 – 32pg), MCHC (320 – 360 g/L), RDW-CV (11 – 14%), PLT (150 – 450 G/L), WBC (4 – 10 G/L). Anemia severity was classified as: mild (Hb 90 – <120 g/L), moderate (Hb 60 – <90 g/L), severe (Hb 30 – <60 g/L), and very severe (Hb < 30 g/L). Daily calibration with control samples ensured a coefficient of variation (CV) < 5%.

Hemoglobin analysis: Hemoglobin quantification was performed using a Bio-Rad Variant II High-Performance Liquid Chromatography (HPLC) system following the Ministry of Health (Vietnam) guidelines. Venous blood samples (2mL, collected in EDTA tubes) were analyzed within 48 hours, with duplicate runs per specimen to ensure precision (inter-assay coefficient of variation < 3%). HPLC served as the reference standard for confirming β-thalassemia, quantifying HbA, HbA$_1$, and HbF according to standardized protocols and internal quality control procedures. Reference ranges were: HbA$_1$ (96 – 98%), HbA$_1$ (0.5 – 3.5%), and HbF (0.1 – 0.5%).[2,12]

Machine Learning Implementation

Algorithms and Techniques

Machine learning models were developed using Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR) algorithms, with preprocessing techniques including Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), Principal Component Analysis (PCA), and Singular Value Decomposition (SVD).[11,13] Experiments were executed on Google Colab using scikit-learn (version 1.2.2) and TensorFlow (version 2.10.0), with hyperparameter optimization performed via grid search (e.g., RF: 100 – 500 trees, DT: max depth 3 – 10, LR: C=0.1 – 10).

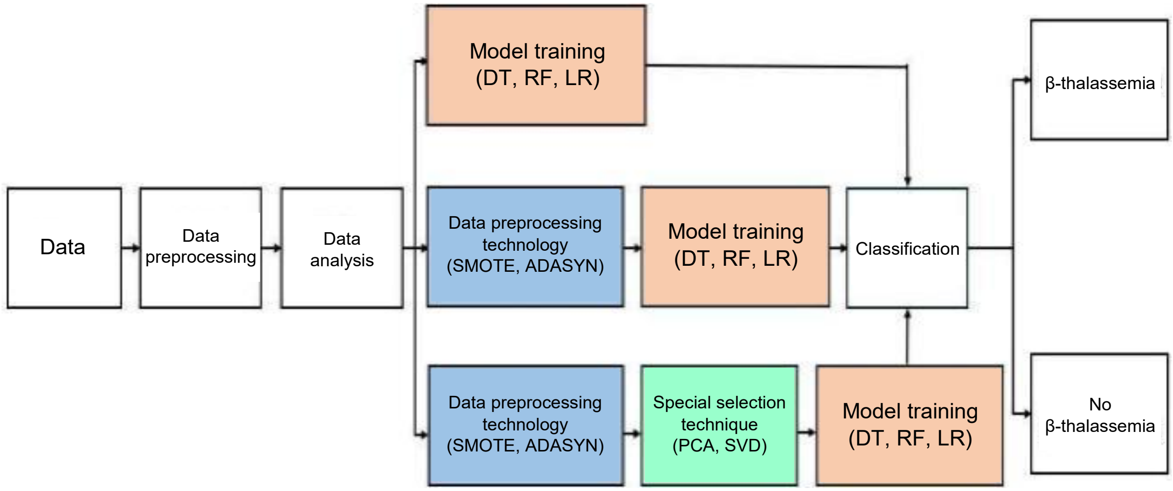

Data
Data preprocessing
Data analysis
Model training (DT, RF, LR)
Data preprocessing technology (SMOTE, ADASYN)
Model training (DT, RF, LR)
Classification
Data preprocessing technology (SMOTE, ADASYN)
Special selection technique (PCA, SVD)
Model training (DT, RF, LR)
β-thalassemia
No β-thalassemia

Figure 1. Machine learning workflow

Data Preprocessing and Model Development

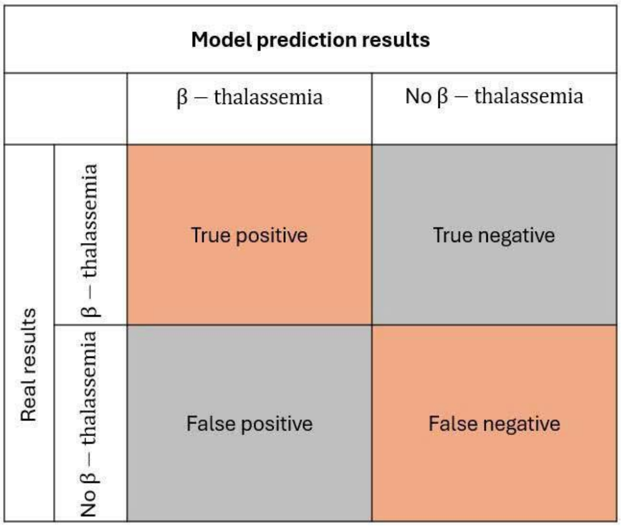
Data were preprocessed using the Label Encoder from scikit-learn to convert categorical variables (e.g., gender, thalassemia status) into numerical values (0 to N-1). Imbalanced class distribution (252 cases vs. 263 controls) was mitigated using SMOTE and ADASYN to oversample the minority class, ensuring balanced representation.[11,13] Missing data were imputed using median values. Feature selection via PCA and SVD reduced dimensionality, retaining the top 5–10 components based on explained variance. The dataset was partitioned into training (80%) and testing (20%) subsets using stratified 5-fold cross-validation to minimize bias. Model performance was evaluated with:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

where TP (True Positive), FP (False Positive), TN (True Negative), and FN (False Negative) are defined in Figure 2.



		Model prediction results	
		β – thalassemia	No β – thalassemia
Real results	β – thalassemia	True positive	True negative
	No β – thalassemia	False positive	False negative

Figure 2. TP/FP/TN/FN definitions

Experimental Design

Initial Experiments:

Experiment 1: Classification using the original imbalanced dataset (252 cases, 263 controls; validation set: 111 samples, 55 β-thalassemia, 56 controls).

Experiment 2: Classification with SMOTE and ADASYN oversampling to address class imbalance.

Experiment 3: Classification with oversampling combined with PCA and SVD for feature selection and dimensionality reduction.

Visual Representation: ML workflow (Figure 1), TP/FP/TN/FN definitions (Figure 2), and experimental design overview (Chart 3) to illustrate methodology.

Follow-up Experiments:

Sample Size Impact: Performance of the optimal model (Random Forest with SMOTE and PCA/SVD) assessed across training sets of 300, 400, and 500 samples, each replicated three times using bootstrap sampling.

Class Balance Impact: Evaluation with a fixed sample size of 400, using stratified sampling across class ratios of 60:40, 50:50, and 40:60 (β-thalassemia:non-β-thalassemia).

Test Set Size Impact: Analysis of model performance with test set proportions ranging from 0.05 to 0.6 in 0.05 increments, using a 10% hold-out validation set for final evaluation.

Statistical Analysis

Statistical analyses were conducted using Microsoft Excel (version 2019) and GraphPad Prism (version 9.0). Results were reported as counts (n) and percentages (%), with model performance metrics including 95% confidence intervals calculated via bootstrap resampling (1,000 iterations). Differences in accuracy between models were assessed using DeLong’s test, with p-values adjusted for multiple comparisons using the Bonferroni correction.

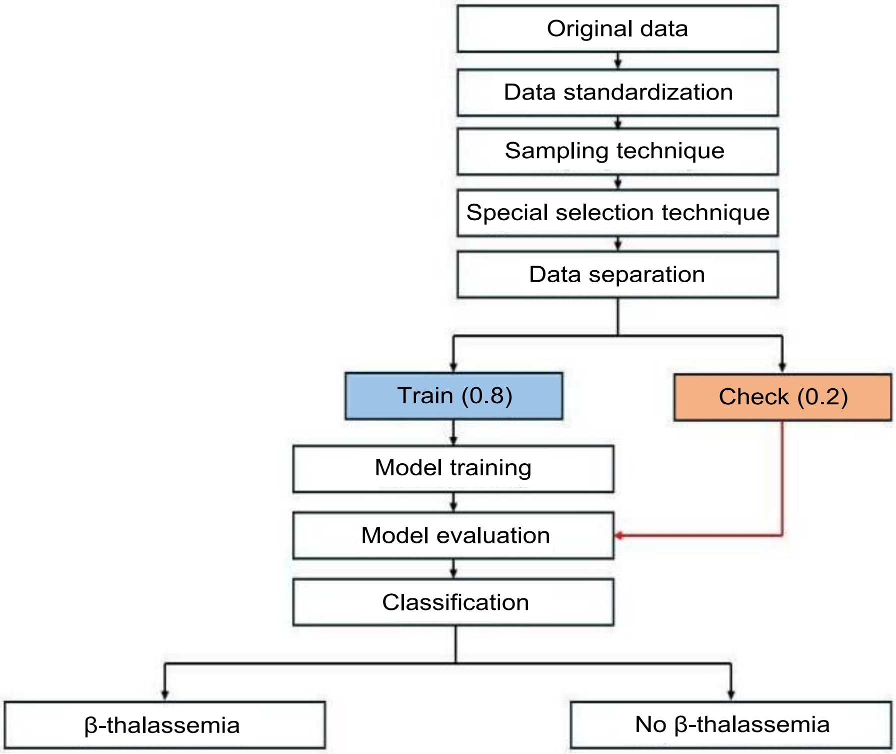


Figure 3. Experimental design

3. Research ethics

The research will respect the confidentiality and anonymity of all subjects. All participants provided with sufficient information about the research subject and are informed about the contents of the research. The research causes no significant harm or threat to participants, physically and emotionally. The research is independent and impartial

III. RESULTS

Model Classification Performance

Machine learning (ML) models (Decision Tree [DT], Random Forest [RF], Logistic Regression [LR]) were evaluated on a validation set of 111 samples (55 β-thalassemia cases, 56 controls) across three experiments, assessing the impact of data preprocessing and parameter optimization. Performance metrics (accuracy, precision, recall, F1-score, AUC) are reported with 95% confidence intervals (CIs).

Experiment 1. Classification Using Original Dataset

Models trained on the original dataset (252 cases, 263 controls) exhibited suboptimal performance (Table 2). LR achieved the highest accuracy (0.77 [95% CI: 0.69 - 0.84]), followed by RF (0.59 [95% CI: 0.51 - 0.67]) and DT (0.57 [95% CI: 0.49 - 0.65]). Confusion matrices revealed high false negatives (e.g., DT: FN = 45; RF: FN = 44), indicating poor sensitivity for β-thalassemia detection.

Table 2. Performance of Machine Learning Models on Original Imbalanced Dataset

Model	Classification	Precision	Recall	F1-score	Accuracy (95% CI)	AUC (95% CI)
DT	Non-β-thalassemia	0.84	0.20	0.33	0.57 (0.49 - 0.65)	0.58 (0.50 - 0.66)
	β-thalassemia	0.54	0.96	0.69		
RF	Non-β-thalassemia	0.92	0.21	0.34	0.59 (0.51 - 0.67)	0.60 (0.52 - 0.68)
	β-thalassemia	0.55	0.98	0.70		
LR	Non-β-thalassemia	0.69	1.00	0.82	0.77 (0.69 - 0.84)	0.77 (0.69 - 0.84)
	β-thalassemia	1.00	0.54	0.70		

*DT, Decision Tree; RF, Random Forest; LR, Logistic Regression. Metrics calculated on a validation set of 111 samples (55 β-thalassemia, 56 controls). Accuracy and AUC are overall model performance; Precision and Recall reported for Non-β-thalassemia and β-thalassemia, respectively; F1-score = 2 * (Precision * Recall) / (Precision + Recall). 95% CIs from bootstrap analysis (1,000 iterations). Confusion matrices: DT (TP = 53, TN = 11, FP = 45, FN = 2), RF (TP = 54, TN = 12, FP = 44, FN = 1), LR (TP = 30, TN = 56, FP = 0, FN = 25)*

Experiment 2. Oversampling with SMOTE and ADASYN

SMOTE oversampling improved performance, with RF achieving the highest accuracy (0.96 [95% CI: 0.92 - 0.99]; AUC, 0.96 [95% CI: 0.93 - 0.98]; F1-score, 0.96). DT and LR reached accuracies of 0.81 (95% CI: 0.74 - 0.87) and 0.90 (95% CI: 0.84 - 0.95), respectively. RF's confusion matrix (TP = 54, TN = 55, FP = 1, FN = 1) showed near-perfect classification. ADASYN yielded variable results: RF and LR accuracies were 0.90 (95% CI: 0.84 - 0.95) and 0.91 (95% CI: 0.85 - 0.96), but DT was poor (0.58 [95% CI: 0.50 - 0.66]). Oversampling results are reported in Supporting Information (Tables S1 and S2).

Experiment 3. Oversampling with Feature Selection

Combining SMOTE with PCA/SVD (Table 3) optimized performance, with RF and DT achieving accuracies of 0.95 (95% CI: 0.91 - 0.98; AUC, 0.96 [95% CI: 0.93 - 0.98]; F1-score, 0.95) and LR at 0.925 (95% CI, 0.88 - 0.96). RF's confusion matrix (TP = 53, TN = 53, FP = 3, FN = 2) indicated high sensitivity (96%) and specificity (95%). ADASYN with PCA/SVD (Table 4) produced the highest DT accuracy (0.97 [95% CI: 0.94 - 0.99]; AUC, 0.97 [95% CI: 0.95 - 0.99]), but RF dropped to 0.85 (95% CI: 0.78 - 0.91), highlighting ADASYN's inconsistency.

Table 3. Performance of Machine Learning Models with SMOTE and PCA/SVD

Model	Classification	Precision	Recall	F1-score	Accuracy (95% CI)	AUC (95% CI)
DT	Non-β-thalassemia	0.93	0.98	0.95	0.95 (0.91 - 0.98)	0.95 (0.91 - 0.98)
	β-thalassemia	0.98	0.93	0.95		
RF	Non-β-thalassemia	0.95	0.96	0.95	0.95 (0.91 - 0.98)	0.96 (0.92 - 0.99)
	β-thalassemia	0.96	0.94	0.95		
LR	Non-β-thalassemia	0.85	1.00	0.92	0.93 (0.88 - 0.96)	0.91 (0.86 - 0.95)
	β-thalassemia	1.00	0.82	0.90		

DT, Decision Tree; RF, Random Forest; LR, Logistic Regression; PCA, Principal Component Analysis; SVD, Singular Value Decomposition. Metrics calculated on a validation set of 111 samples (55 β-thalassemia, 56 controls). Precision, Recall, and F1-score reported for Non-β-thalassemia (upper row) and β-thalassemia (lower row) per model. 95% CIs from bootstrap analysis (1,000 iterations). Confusion matrices: DT (TP = 54, TN = 55, FP = 1, FN = 1), RF (TP = 53, TN = 53, FP = 3, FN = 2), LR (TP = 45, TN = 56, FP = 0, FN = 10)

Table 4. Performance of Machine Learning Models with ADASYN and PCA/SVD

Model	Classification	Precision	Recall	F1-score	Accuracy (95% CI)	AUC (95% CI)
DT	Non-β-thalassemia	0.94	1.00	0.97	0.97 (0.94 - 0.99)	0.97 (0.94 - 0.99)
	β-thalassemia	1.00	0.94	0.97		
RF	Non-β-thalassemia	0.97	0.71	0.82	0.85 (0.78 - 0.91)	0.84 (0.77 - 0.90)
	β-thalassemia	0.77	0.98	0.86		
LR	Non-β-thalassemia	0.85	1.00	0.92	0.92 (0.88 - 0.96)	0.91 (0.86 - 0.95)
	β-thalassemia	1.00	0.82	0.90		

DT, Decision Tree; RF, Random Forest; LR, Logistic Regression; PCA, Principal Component Analysis; SVD, Singular Value Decomposition. Metrics calculated on a validation set of 111 samples (55 β-thalassemia, 56 controls). Precision, Recall, and F1-score reported for Non-β-thalassemia (upper row) and β-thalassemia (lower row) per model. 95% CIs from bootstrap analysis (1,000 iterations). Confusion matrices: DT (TP = 55, TN = 54, FP = 1, FN = 1), RF (TP = 54, TN = 40, FP = 2, FN = 15), LR (TP = 45, TN = 56, FP = 0, FN = 10)

Parameter Optimization

Using RF with SMOTE and PCA/SVD, training set size impacted performance (Table 5). A 500-sample set yielded the highest accuracy (0.95 [95% CI: 0.91 - 0.98]), declining to 0.92 (95% CI, 0.87-0.96) at 400 samples and 0.90 (95% CI: 0.84 - 0.94) at 300 samples. Class balance (Table 6) showed a 60:40 ratio as optimal (0.94 [95% CI: 0.90 - 0.97]), with 50:50 and 40:60 at 0.92 (95% CI: 0.87 - 0.96). Test set size results are reported in Supporting Information (Table S3). The AUC performance of the Random Forest (RF) model with SMOTE+PCA/SVD across varying test set sizes is shown in Chart 1.

Table 5. Performance of Random Forest with SMOTE and PCA/SVD Across Training Set Sizes

Training Set Size	Classification	Precision	Recall	F1-score	Accuracy (95% CI)	AUC (95% CI)
500 samples	Non-β-thalassemia	0.95	0.96	0.95	0.95 (0.91 - 0.98)	0.96 (0.92 - 0.99)
	β-thalassemia	0.96	0.94	0.95		
400 samples	Non-β-thalassemia	0.94	0.89	0.91	0.92 (0.87 - 0.96)	0.92 (0.87 - 0.96)
	β-thalassemia	0.89	0.94	0.91		
300 samples	Non-β-thalassemia	0.81	1.00	0.89	0.90 (0.84 - 0.94)	0.88 (0.82 - 0.93)
	β-thalassemia	1.00	0.76	0.86		

Metrics calculated using Random Forest with SMOTE, PCA, and SVD on a validation set of 111 samples (55 β-thalassemia, 56 controls). Precision, Recall, and F1-score reported for Non-β-thalassemia (upper row) and β-thalassemia (lower row). 95% CIs from bootstrap analysis (1,000 iterations). Confusion matrices: 500 samples (TP = 53, TN = 53, FP = 3, FN = 2), 400 samples (TP = 49, TN = 50, FP = 5, FN = 7), 300 samples (TP = 42, TN = 56, FP = 0, FN = 13)

Table 6. Performance of Random Forest with SMOTE and PCA/SVD Across Class Ratios

Class Ratio	Classification	Precision	Recall	F1-score	Accuracy (95% CI)	AUC (95% CI)
60:40	Non-β-thalassemia	0.90	0.98	0.94	0.94 (0.90 - 0.97)	0.94 (0.90 - 0.97)
	β-thalassemia	0.98	0.88	0.93		
50:50	Non-β-thalassemia	0.90	0.95	0.92	0.92 (0.87 - 0.96)	0.92 (0.87 - 0.96)
	β-thalassemia	0.94	0.89	0.91		
40:60	Non-β-thalassemia	0.90	0.95	0.92	0.92 (0.87 - 0.96)	0.92 (0.87 - 0.96)
	β-thalassemia	0.94	0.89	0.91		

Metrics calculated using Random Forest with SMOTE, PCA, and SVD on a validation set of 111 samples (55 β-thalassemia, 56 controls). Precision, Recall, and F1-score reported for Non-β-thalassemia (upper row) and β-thalassemia (lower row). 95% CIs from bootstrap analysis (1,000 iterations). Confusion matrices: 60:40 (TP = 54, TN = 55, FP = 1, FN = 1), 50:50 (TP = 52, TN = 53, FP = 3, FN = 3), 40:60 (TP = 52, TN = 53, FP = 3, FN = 3)

The RF model was tested with test sizes from 0.05 to 0.6 on a 500-sample training set. Results are shown in Table S3 and Chart 1.

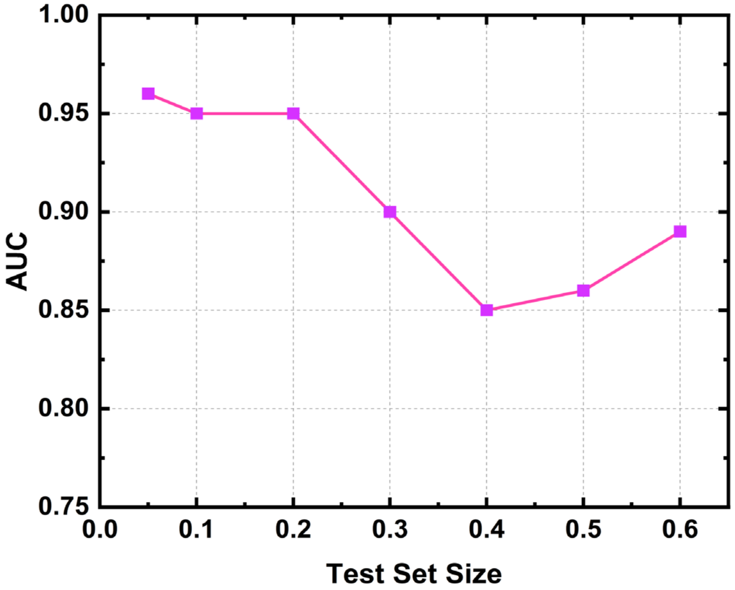


Chart 1. AUC vs Test Set Size for the Random Forest Model (SMOTE + PCA)

The receiver operating performance of the Random Forest (RF) model trained with SMOTE and PCA is shown across varying test set proportions (0.05 – 0.6). The highest AUC values (0.95 – 0.96) were observed for smaller test sizes (0.05 – 0.2), indicating model stability and optimal generalization under moderate data partitioning. AUC decreased when the test proportion exceeded 0.3, suggesting sensitivity to limited training data.[8]

IV. DISCUSSION

This study evaluated machine learning (ML) models for β-thalassemia screening using complete blood count (CBC) data from 515 students in Lai Châu province, Vietnam (49% β-thalassemia, 51% controls), excluding iron-deficiency anemia cases.[3] Three experiments explored the effects of data preprocessing, feature extraction, and parameter optimization, demonstrating that ML can serve as a cost-effective and non-invasive screening tool for β-thalassemia in resource-limited settings.

Initial experiments on the original imbalanced dataset revealed modest accuracies (Logistic Regression [LR]: 0.77 [95% CI: 0.69 - 0.84], Random Forest [RF]: 0.59 [95% CI: 0.51 - 0.67], Decision Tree [DT]: 0.57 [95% CI: 0.49 - 0.65]), driven by high false negatives (e.g., DT: FN = 45), underscoring the challenge of class imbalance in hematologic ML tasks.[4] These findings are consistent with prior studies, such as Rustam et al. (2022), who reported accuracies of 0.90 – 0.91 on balanced datasets, and Saleem et al. (2023), who achieved 0.83 – 0.88 on imbalanced data, reinforcing the need for data-balancing techniques.[8,10]

Oversampling with SMOTE and ADASYN, combined with PCA/SVD, significantly improved model performance. SMOTE with PCA/SVD yielded optimal results (RF: 0.95 [95% CI: 0.91 - 0.98], DT: 0.95 [95% CI: 0.91 - 0.98], LR: 0.93 [95% CI: 0.88 - 0.96]; AUC ≈ 0.96; F1-score ≈ 0.95), with RF demonstrating high sensitivity (98%) and specificity (96%) (TP = 53, TN =

53, FP = 3, FN = 2).[11] ADASYN with PCA/SVD excelled with DT (0.97 [95% CI: 0.94 - 0.99], AUC 0.97), but RF performance declined to 0.85 [95% CI: 0.78 - 0.91], reflecting model-specific responses to synthetic data generation, as noted by Chawla et al. (2002).[11] These outcomes align with DeepThal (2022) and Christensen et al. (2025), who reported AUCs > 0.95, despite our study's smaller, regionally focused cohort.[14]

Parameter optimization further refined results. A 500-sample training set maximized RF accuracy (0.95 [95% CI: 0.91 - 0.98]), declining to 0.90 [95% CI: 0.84 - 0.94] at 300 samples, supporting Li et al.'s (2021) emphasis on dataset size.[12] A 60:40 class ratio optimized accuracy (0.94 [95% CI: 0.90 - 0.97]), suggesting mild imbalance mirrors real-world prevalence, consistent with Batista et al. (2004).[15] Test set size analysis (Table S3) showed peak AUCs (0.95 – 0.96) at 0.05 – 0.2, declining beyond 0.3, indicating sensitivity to training data availability.

Compared to traditional hematologic indices such as Mentzer (MCV/RBC, sensitivity 74 – 90%) and Shine & Lal (MCV^2 × MCH × 0.01), the optimized RF model achieved markedly higher diagnostic accuracy (sensitivity 98%, specificity 96%), substantially reducing false negatives-an essential advantage for early carrier detection and public health screening.[16,17] Integrating such ML tools into Vietnam's existing CBC-based screening workflows could enhance accessibility, particularly in remote or low-resource areas where molecular testing remains unavailable.

Nevertheless, several limitations should be acknowledged. This study relied solely on HPLC as the confirmatory method without genetic validation. Although HPLC is considered sufficient by WHO, TIF, and ARUP guidelines, it may not detect silent or compound mutations.[1-3] Furthermore, the single-center design and relatively small sample size may limit generalizability. Future multi-center studies incorporating genetic confirmation and larger, demographically diverse cohorts are needed to verify the scalability and clinical utility of ML-based thalassemia screening in Vietnam.

V. CONCLUSION

This study demonstrated that machine learning (ML) models, particularly Random Forest and Decision Tree algorithms enhanced with SMOTE and PCA/SVD preprocessing, can accurately screen for β-thalassemia using complete blood count (CBC) parameters alone. The approach achieved high diagnostic performance while maintaining affordability and simplicity, supporting its potential integration into routine hematology screening in resource-limited settings.

However, as this single-center study was based on a relatively small dataset without genetic confirmation, further validation across multiple centers and inclusion of molecular data are warranted to confirm the generalizability and clinical utility of ML-based β-thalassemia screening in Vietnam.

Acknowledgments

This study was supported by the institutional scientific research project of Hanoi Medical University (Grant No. 282/QD-DHYHN).

REFERENCES

1. Thalassaemia International Federation. *Guidelines for the Management of Transfusion-Dependent Thalassaemia (TDT).* 3rd ed. Nicosia, Cyprus: Thalassaemia International Federation; 2022.

2. World Health Organization. *Management of Haemoglobin Disorders: Report of a Joint WHO–TIF Meeting.* Geneva, Switzerland:

World Health Organization; 2008.

3. ARUP Consult. Thalassemias-Choose the Right Test. Updated 2025. Accessed October 13, 2025. https://arupconsult.com/content/thalassemias

4. Weatherall DJ. The inherited disorders of haemoglobin: an increasingly neglected global health burden. *Indian J Med Res.* 2011;134(4):493-497. doi:10.4103/0971-5916.90987

5. Taher AT, Weatherall DJ, Cappellini MD, et al. Thalassaemia. *Lancet.* 2018;391(10116):155-167. doi:10.1016/S0140-6736(17)31822-6

6. Galanello R, Origa R. Beta-thalassemia. *Orphanet J Rare Dis.* 2010;5:11. doi:10.1186/1750-1172-5-11

7. Modell B, Darlison M. Global epidemiology of haemoglobin disorders and derived service indicators. *Bull World Health Organ.* 2008;86(6):480-487. doi:10.2471/BLT.06.036673

8. Saleem M, Ali S, Khan MA, et al. Automated detection of thalassemia using machine learning techniques. *Comput Methods Programs Biomed.* 2023;231:107407. doi:10.1016/j.cmpb.2023.107407

9. Nguyen Ba Tung, Tran Danh Cuong, Nguyen Thi Trang, et al. Research on the application of artificial intelligence in prenatal screening for thalassemia. *Vietnam Medical Journal*. 2023;526(2). doi:10.51298/vmj.v526i2.5590

10. Rustam F, Ashraf I, Jabbar S, et al. Prediction of β-thalassemia carriers using complete blood count features. *Sci Rep.* 2022;12(1):19999. doi:10.1038/s41598-022-22011-8

11. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res.* 2002;16:321–357. doi:10.1613/jair.953

12. Devkota BP. Hemoglobin Electrophoresis-Reference Range. *Medscape.* Updated July 2, 2025. Accessed October 13, 2025. https://emedicine.medscape.com/article/2085637-overview

13. He H, Bai Y, Garcia EA, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks.* Piscataway, NJ: IEEE; 2008:1322-1328. doi:10.1109/IJCNN.2008.4633969

14. Christensen F, Kılıç DK, Nielsen IE, et al. Classification of α-thalassemia data using machine learning models. *Comput Methods Programs Biomed.* 2025;260:108581. doi:10.1016/j.cmpb.2024.108581

15. World Health Organization. *Hemoglobin Concentrations for the Diagnosis of Anemia and Assessment of Severity.* Geneva, Switzerland: World Health Organization; 2011.

16. AlQarni AM, Althumairi A, Alkaltham NK, et al. Diagnostic test performance of the Mentzer index in evaluating Saudi children with microcytosis. *Front Med (Lausanne).* 2024;11:1361805. doi:10.3389/fmed.2024.1361805

17. Shah TP, Shrestha A, Agrawal JP, et al. Role of Mentzer Index for differential diagnosis of iron deficiency anaemia and beta thalassemia trait. *J Nepal Health Res Counc.* 2023;21(1):99-102. doi:10.33314/jnhrc.v21i1.4479