



Journal of Associated
Medical Sciences

Scopus Indexed & Thai-Journal Citation Index Centre (TCI)

Journal of Associated Medical Sciences

Journal homepage: <https://www.tci-thaijo.org/index.php/bulletinAMS/index>



Machine learning with SMOTE-NC on routine hematology parameters for early identification of Hb E trait

Chollanot Kaset^{1,2}, Tanita Rerkchalerm¹, Tanatorn Tanantong^{3,4}, Sirinart Chomean^{1,2*}

¹Graduate Program in Medical Technology, Faculty of Allied Health Sciences, Thammasat University, Pathum Thani, Thailand.

²Thammasat Research Unit in Medical Technology and Precision Medicine Innovation, Faculty of Allied Health Sciences, Thammasat University, Pathum Thani, Thailand.

³Department of Computer Science, Faculty of Science and Technology, Thammasat University, Pathum Thani, Thailand.

⁴Thammasat Research Unit in Data Innovation and Artificial Intelligence, Thammasat University, Pathum Thani, Thailand.

ARTICLE INFO

ORCID of all authors

Chollanot Kaset: <https://orcid.org/0000-0002-2110-6918>

Tanita Rerkchalerm: <https://orcid.org/0009-0005-6019-1907>

Tanatorn Tanantong: <https://orcid.org/0000-0002-8927-2365>

Sirinart Chomean: <https://orcid.org/0000-0003-2310-2934>

Correspondence

Sirinart Chomean

Graduate Program in Medical Technology, Faculty of Allied Health Sciences, Thammasat University, Pathum Thani, Thailand.

Email: sirinat.c@allied.tu.ac.th

Article history

Received 8 April 2026

Accepted as revised 24 August 2026

Available online 2 September 2026

Keywords

Thalassemia, Hb E trait, support vector machine, SMOTE-NC, CBC

ABSTRACT

Background: The hemoglobin E (Hb E) trait, which is highly prevalent in Southeast Asia and may co-occur with α -thalassemia, often exhibits heterogeneous complete blood count profiles; notably, some individuals with the Hb E trait have near-normal MCV and MCH values, increasing the risk of misclassification during routine hematological screening.

Objective: This study aimed to develop and externally validate an interpretable machine-learning model to screen for Hb E trait using routine Complete Blood Count (CBC) data.

Materials and methods: We analyzed 5,787 standardized CBC records using two processing pipelines: a standard dataset and an SMOTE-NC oversampled dataset. We selected features using ReliefF and evaluated nine machine-learning algorithms. We interpreted the optimal model using SHAP values and validated it on an independent external cohort (n=625).

Results: Discrimination for Hb E trait was moderate-to-good and consistent between internal and external evaluation. Erythrocyte indices-Z-MCV, Z-MCH, and Z-RBC-were the most discriminative features. Among nine algorithms, a support vector machine gave the best combination of sensitivity and discrimination (internal AUC 0.881, sensitivity 70.66%, specificity 87.92%; external AUC 0.879, sensitivity 72.38%, specificity 87.39%). With the decision threshold tuned for screening, external sensitivity rose to approximately 91% with negative predictive value above 94%. SHAP analysis confirmed the model's reliance on biologically plausible drivers (low MCV/MCH with a relatively preserved RBC count).

Conclusion: An explainable machine-learning model utilizing routine CBC parameters offers a viable, low-cost pre-screening tool for Hb E trait. Internal and external performance are concordant, and with local calibration and threshold tuning, the model triages patients for confirmatory testing and definitive genotyping rather than replacing them, supporting optimized diagnostic workflows in endemic regions.

Introduction

Hemoglobin E (Hb E) is one of the most prevalent structural hemoglobin variants worldwide and is especially common in Southeast Asia, including Thailand, Cambodia, Laos, Myanmar, Sri Lanka, and parts of India.¹⁻⁵ The World Health Organization estimates that approximately 7% of the global population carries clinically relevant hemoglobin variants, with Hb E representing a major contributor in Southeast Asian populations.^{1,4} In Thailand, the prevalence of Hb E trait ranges from 5-10%, making it a major public health concern and a key target of national thalassemia prevention programs.^{2,3,6} Individuals with heterozygous Hb E are typically asymptomatic and clinically healthy. Routine complete blood count (CBC) testing may reveal mild microcytosis or, in some cases, near-normal red cell indices, complicating early recognition during population screening.⁷⁻⁹

The clinical significance of Hb E increases substantially when it is co-inherited with other hemoglobinopathies, particularly α -thalassemia or β -thalassemia. Compound genotypes such as Hb E/ β -thalassemia are among the most common causes of clinically significant thalassemia syndromes in Thailand and neighboring countries.^{3,6} In addition, homozygous Hb E or Hb E trait co-inherited with α -thalassemia may produce heterogeneous hematologic phenotypes, with variable effects on MCV, MCH, and RBC count, further complicating diagnostic interpretation.⁸ Accurate identification of Hb E trait and its coinheritance patterns is therefore essential for genetic counseling,

prenatal screening, and disease prevention strategies in high-prevalence regions.

In Thailand, first-line thalassemia screening relies on a combination of hematological indices-particularly mean corpuscular volume (MCV) and mean corpuscular hemoglobin (MCH)-together with dichlorophenol indophenol precipitation (DCIP) testing, osmotic fragility (OF) testing, and hemoglobin typing using high-performance liquid chromatography (HPLC) or capillary electrophoresis.^{2,7,9} While these approaches are effective for initial screening, definitive classification of α -thalassemia still requires PCR-based or other molecular genotyping.⁶ This stepwise diagnostic workflow is labor-intensive and subject to inter-observer variability, especially in cases with overlapping hematologic features or borderline indices.^{7,8} As screening volumes increase, scalable and interpretable automated tools are needed to support early triage and reduce unnecessary confirmatory testing.

In recent years, artificial intelligence (AI) and machine learning (ML) approaches have gained increasing attention in clinical hematology. Numerous studies have applied algorithms such as support vector machines (SVM), artificial neural networks (ANN), random forest (RF), decision trees, k-nearest neighbors, and gradient boosting to classify thalassemia traits using routine CBC parameters.¹⁰⁻¹⁵ For example, ThalPred employed SVM-based modeling to differentiate thalassemia traits from normal and iron-deficient states with high accuracy using a limited set of red cell indices.¹⁰ Subsequent studies from South Asia

and the Middle East reported comparable performance using ensemble and tree-based methods for β -thalassemia carrier detection.¹¹⁻¹³ However, most prior work focused on binary classification problems or β -thalassemia-dominant cohorts, with limited attention to the diagnostically complex subgroup of Hb E trait and its coinheritance with α -thalassemia.

A further challenge in ML-based hematologic modeling is class imbalance. In real-world screening datasets, cases of Hb E trait with or without α -thalassemia occur far less frequently than normal or non-thalassemic samples. Models trained on imbalanced data tend to favor majority classes, resulting in reduced sensitivity for clinically important minority groups.^{14,16} Oversampling techniques such as the Synthetic Minority Oversampling Technique (SMOTE) have been widely adopted to mitigate this issue.¹⁶⁻⁸ When datasets include both continuous and categorical variables, SMOTE-NC provides an appropriate extension that preserves the structure of mixed-type hematological data.¹⁷

Open-source Python libraries-including scikit-learn, XGBoost, CatBoost, and imbalanced-learn-have become standard tools for developing machine-learning models in biomedical research, providing transparent, script-based workflows for preprocessing, feature selection, model training, and performance evaluation that support reproducibility and interpretability. Such tools have been applied to anemia and hemoglobinopathy classification with results validated against expert interpretation.

Despite these advances, few studies have specifically addressed automated classification of Hb E trait and Hb E trait with or without α -thalassemia using routine hematology parameters. This gap is particularly relevant in Southeast Asia, where Hb E is endemic, and coinheritance patterns are common.^{3,6} The present study aims to develop and validate machine-learning models, implemented in Python, to identify Hb E trait and its coinheritance patterns from routine hematological data.

By integrating SMOTE-NC for class balancing, benchmarking nine algorithms, and using explainable

AI (SHAP) for interpretation, this work seeks to improve early screening accuracy, reduce diagnostic burden, and support effective thalassemia prevention strategies in high-prevalence settings.

Materials and methods

Data collection and study design

All hematological data used in this research were retrospective and fully anonymized prior to analysis. Investigators did not access any personally identifiable information. Because the study was retrospective and used de-identified routine laboratory data, the ethics committee waived the requirement for informed consent. The study involved no direct patient contact and posed no additional risk to participants.

This study aimed to develop and validate machine-learning models to screen for the Hb E trait and Hb E trait with or without α -thalassemia using routine CBC parameters, a task of particular relevance in Thailand and Southeast Asia, where Hb E is endemic.^{2,3,19} Hematological data were retrospectively collected over a five-year period and curated in Microsoft Excel. The dataset comprised demographic variables (sex and age), CBC results, and hemoglobin typing outcomes. All personal identifiers were removed and replaced with newly assigned study codes.

The initial dataset comprised 5,787 records. Data-cleaning procedures excluded incomplete entries; records with more than 10% missing values were removed from further analysis, leaving 5,778 records for modeling. Hemoglobin typing results, interpreted according to the national laboratory diagnostic manual issued by the Department of Medical Sciences, Thailand, served as the gold standard for labeling and class assignment.²

The overall workflow and analytical pipeline were schematically illustrated using Eraser.io (<https://app.eraser.io/>, © 2025 Eraser Labs, Inc.) to generate reproducible diagrams representing data preprocessing, model development, validation, and interpretation (Figure 1).

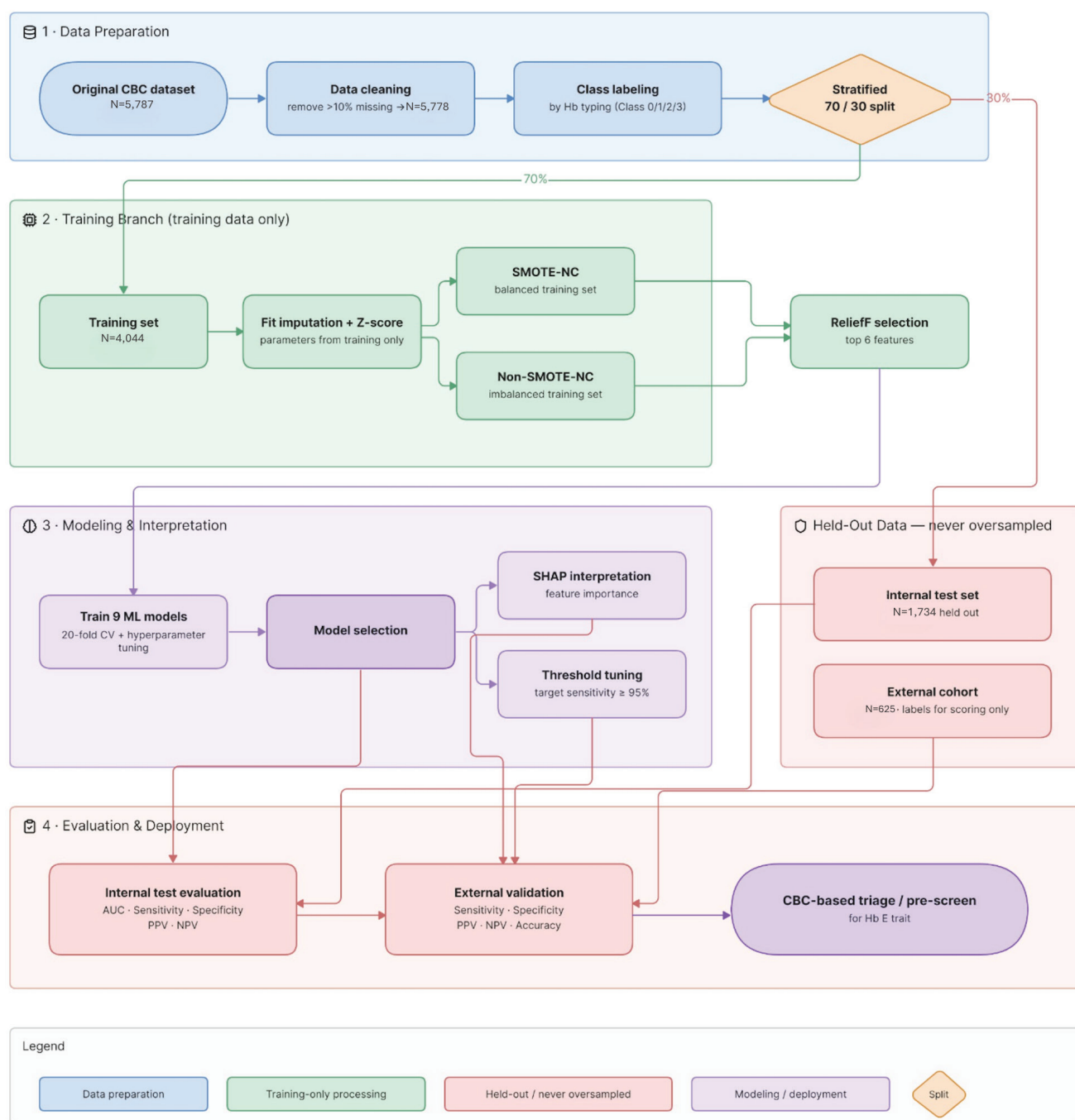


Figure 1. Workflow for developing and validating machine-learning models to classify Hb E trait, with or without α -thalassemia (Class 2).

Data grouping

Based on hemoglobin typing results and clinical relevance, samples were categorized into four diagnostic classes. Because Class 0 and Class 1 both display a normal hemoglobin-typing pattern, the operational criteria used to separate them are defined explicitly in Table 1, Class 0 comprises individuals with a normal typing pattern and normal red-cell indices (no clinically significant thalassemia), whereas Class 1

comprises individuals with a normal typing pattern who may carry α -thalassemia and can therefore present with reduced MCV/MCH despite the absence of an abnormal hemoglobin fraction, since α -thalassemia produces no abnormal band on typing.²⁰ Class 2 corresponds to Hb E trait with or without co-inherited α -thalassemia, in which mild microcytosis coexists with a relatively preserved RBC count^[1,19].

Table 1. Definition of the four diagnostic classes by hemoglobin-typing pattern and hematologic/genetic characteristics.

Class	Group	Hemoglobin typing	Hematologic / genetic characteristics
0	Normal/non-clinically significant	Normal (A2A; HbA2 within reference range)	Normal red-cell indices; no clinically significant thalassemia.
1	Normal typing $\pm\alpha$ -thalassemia	Normal (A2A)	May carry α -thalassemia (α^+ or α^0); microcytosis/hypochromia can occur with normal typing. Definitive α -status requires molecular testing.
2	Hb E trait $\pm\alpha$ -thalassemia	Hb E trait (EA; HbE ~25-30%, lower with α_0)	Mild microcytosis with relatively preserved RBC count; screening target. Co-inherited α -thalassemia further lowers MCV/MCH and HbE fraction.
3	Other hemoglobinopathies	Abnormal (Homozygous Hb E, Hb H, Hb CS, β -thalassemia $\pm$ α -thalassemia)	Heterogeneous, often more pronounced red-cell abnormalities.

Data preprocessing, data splitting, and class balancing

To eliminate any risk of information leakage between preprocessing and evaluation, the analysis followed a strict order. First, the original cleaned dataset (N=5,778) was partitioned using a stratified 70/30 split into a training set (N=4,044) and an internal test set (N=1,734), preserving the class distribution. Only afterward were the remaining steps applied. Median/mode imputation and Z-score standardization were fit on the training set only and applied unchanged to the internal test set and external cohort. Class balancing with SMOTE-NC was applied exclusively to the 70% training partition. SMOTE-NC extends SMOTE to mixed continuous-categorical data and suits CBC data combined with sex. The internal test set and external cohort were never oversampled and contained no synthetic observations, so all test and external metrics reflect real class prevalence; confining oversampling to training data follows recommended practice, because oversampling before partitioning can introduce optimistic bias.¹⁰⁻¹² SMOTE-NC was implemented in Python (imbalanced-learn), oversampling minority classes to the majority training-class size with a fixed random seed. A parallel non-SMOTE-NC pipeline used the identical split and preprocessing but trained on the original imbalanced training partition (Figure 1).

Feature selection

Feature ranking used the ReliefF algorithm, fit on each pipeline's training partition, and the top six predictors were retained. Restricting feature selection to training data preserved the independence of the internal test and external sets.

Model construction and evaluation

All analyses were performed in Python within Google Colab (NVIDIA A100 GPU) using scikit-learn, XGBoost, CatBoost, imbalanced-learn (SMOTE-NC),

skrebate (ReliefF), and SHAP. Nine machine-learning classifiers were implemented with the following hyperparameters: Decision Tree (minimum leaf size=10; maximum depth=100); Logistic Regression (L2 regularization; C=1); Random Forest (100 trees; bootstrap sampling); Support Vector Machine (radial-basis-function kernel; C=1; probability estimates enabled); Gradient Boosting Classifier (100 trees; learning rate=0.1; maximum depth=3); XGBoost (100 trees; learning rate=0.3; L2 regularization λ =1; maximum depth=6); CatBoost (100 trees; learning rate=0.3; L2 leaf regularization=3; depth=6); Gaussian Naïve Bayes; and a Neural Network (multilayer perceptron with two hidden layers of 10 and 6 units; ReLU activation; Adam optimizer; α =0.0001; 200 iterations). A fixed random seed was applied throughout for reproducibility. The model previously labeled "Scikit-learn" is the Gradient Boosting Classifier and is named accordingly throughout.

Models were trained as four-class classifiers (Class 0/1/2/3), and all diagnostic metrics were reported for Class 2 (Hb E trait) under a one-versus-rest scheme. We compared models and assessed final performance on the internal test set using AUC, classification accuracy, F1-score, precision, recall, Matthews correlation coefficient (MCC), specificity, and confusion matrices; model selection prioritized sensitivity (recall), F1-score, MCC, and AUC. Five-fold stratified cross-validation on the training set was used solely to select the sensitivity-oriented decision threshold (below), without using test-set data.

Sensitivity-oriented threshold tuning

Because a screening tool must minimize false negatives, the probability threshold applied to P (Class 2) was selected to reach a target sensitivity of $\geq 95\%$ using out-of-fold cross-validated predictions on the training set only; the fixed threshold was then applied unchanged to the internal test set and external cohort, adjusting the operating point without retraining and without using test data.

Model prediction and external validation

The selected classifiers were evaluated on 1) the internal holdout test set (30%), and 2) an independent external cohort of 625 CBC profiles. The external cohort carried its own hemoglobin-typing-based class label, used solely as ground truth for scoring; it was never used for training, scaling, oversampling, or feature selection. Metrics comprised sensitivity, specificity, PPV, NPV, accuracy, and AUC. SHAP analysis was applied to the selected model for transparent interpretation.

Results

Baseline class distribution and data partition

After cleaning, 5,778 records were available. The dataset was markedly imbalanced: Class 0: 3,222 (55.8%), Class 1: 720 (12.5%), Class 2: 1,363 (23.6%), and Class 3: 473 (8.2%). Stratified partitioning produced 4,044 training and 1,734 internal-test records; the external cohort comprised 625 profiles, of which 181 (29.0%) were Class 2 (Table 2).

Table 2. Baseline sample size and proportion for each class (pre-SMOTE-NC), with the stratified partition.

Class	Total N (%)	Training (N)	Internal test (N)
0: Normal / non-significant	3,222 (55.8%)	2,255	967
1: Normal $\pm$ α -thalassemia	720 (12.5%)	504	216
2: Hb E trait $\pm$ α -thalassemia	1,363 (23.6%)	954	409
3: Other hemoglobinopathies	473 (8.2%)	331	142
Total	5,778 (100%)	4,044	1,734

Feature selection using ReliefF

ReliefF ranking, computed on each pipeline's training partition, prioritized erythrocyte indices. In the SMOTE-NC pipeline, the top six features were MCV, MCH, MCHC, RDW, neutrophil percentage (NEU), and RBC, with MCV and MCH dominant; the non-SMOTE-NC pipeline retained a comparable erythrocyte-dominated set.

Model performance comparison (Internal test)

On the internal test set, discrimination for Class 2 was moderate to good in both the non-SMOTE-NC and SMOTE-NC pipelines, with no model approaching the previously reported AUC of $\sim$ 0.99. In the non-SMOTE-NC pipeline, AUC ranged from 0.816 (Decision Tree) to 0.872, with Gradient Boosting (0.872), the support vector machine (SVM, RBF kernel; 0.870), CatBoost (0.869), logistic regression (0.869), and the neural network (0.868) clustered at the top; Class 2 sensitivity ranged from about 57% to 73%, highest for logistic

regression (72.62%), Naïve Bayes (70.90%), and the SVM (70.17%) (Table 3 and Figure 2A). In the SMOTE-NC pipeline, AUC ranged from 0.790 (Decision Tree) to 0.881, and the strongest models by the screening-priority metrics were the SVM (AUC 0.881; sensitivity 70.66%), Gradient Boosting (AUC 0.878), the neural network (AUC 0.876), and CatBoost (AUC 0.869) (Table 3 and Figure 2B). Across both pipelines, the SVM provided the most consistent combination of discrimination and sensitivity while maintaining high specificity, and was therefore selected as the primary model; in the SMOTE-NC pipeline it achieved AUC 0.881, sensitivity 70.66%, specificity 87.92%, PPV 64.37%, NPV 90.66%, and accuracy 83.85% at the default operating point (Table 3). Confusion-matrix analysis (Figure 3) showed that residual Class 2 errors were distributed mainly to Class 1 and Class 0-both normal-typing groups-reflecting genuine phenotypic overlap rather than model instability.

Table 3. Internal-test performance for Class 2 (Hb E trait) across nine algorithms, comparing the non-SMOTE-NC and SMOTE-NC pipelines (one-vs-rest).

Model	non-SMOTE-NC				SMOTE-NC			
	AUC	Sens%	Spec%	Acc%	AUC	Sens%	Spec%	Acc%
Decision Tree	0.816	57.21	88.53	81.14	0.790	47.43	89.43	79.53
Logistic Regression	0.869	72.62	82.72	80.33	0.847	66.01	84.08	79.82
Random Forest	0.865	59.90	88.08	81.43	0.866	55.75	90.34	82.18
SVM (RBF)	0.870	70.17	85.74	82.06	0.881	70.66	87.92	83.85
Gradient Boosting	0.872	64.55	87.47	82.06	0.878	58.92	90.11	82.76
XGBoost	0.849	56.97	88.15	80.80	0.861	53.79	90.79	82.06
CatBoost	0.869	63.57	87.55	81.89	0.869	57.21	90.11	82.35
Naïve Bayes	0.849	70.90	84.30	81.14	0.841	70.66	84.91	81.55
Neural Network	0.868	68.95	86.19	82.12	0.876	61.37	89.81	83.10

Note: Values are AUC and percentages; full metrics (PPV, NPV, F1, MCC) are given in the analysis output.

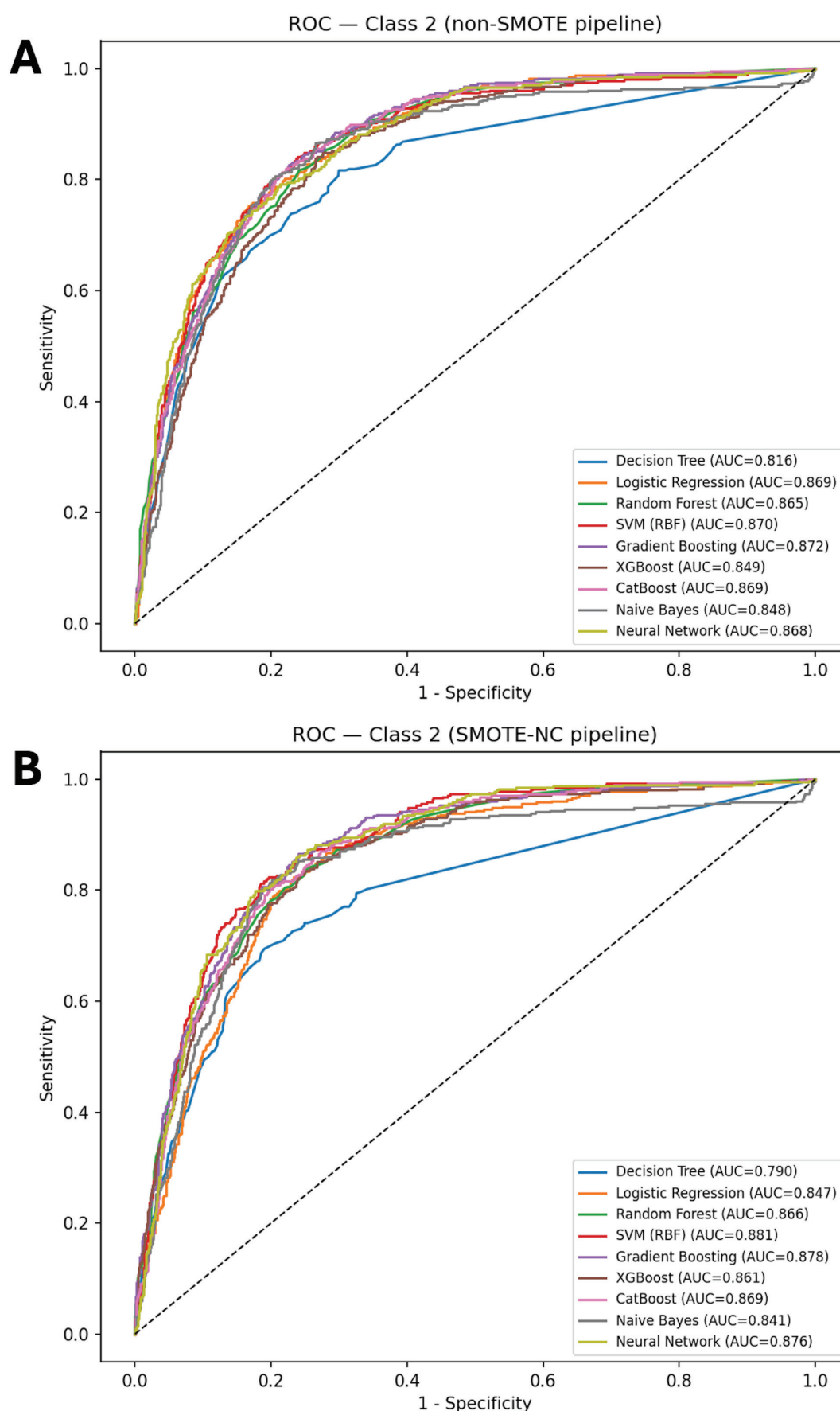


Figure 2. Receiver-operating-characteristic (ROC) curves for Class 2 (Hb E trait with or without α -thalassemia) identification across the nine algorithms, evaluated on the untouched internal test set. A: Non-SMOTE-NC pipeline, B: SMOTE-NC pipeline. Areas under the curve (AUC) are shown in each legend. N=1,734.

Confusion matrices — SMOTE-NC pipeline (internal test)

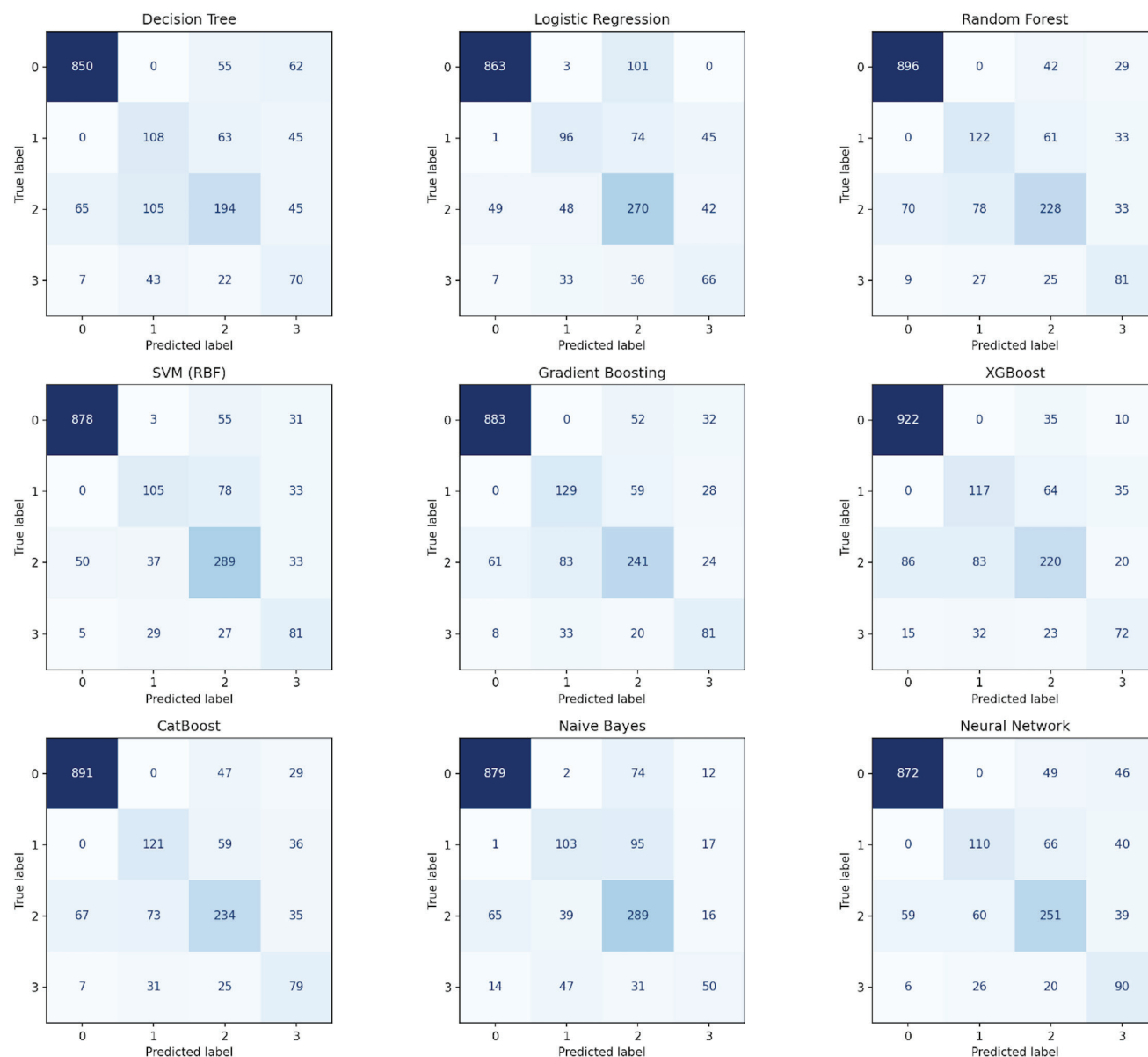


Figure 3. Confusion matrices for the nine algorithms on the internal test set (SMOTENC pipeline). Rows are true classes and columns predicted classes (Class 0: normal/nonsignificant, Class 1: normal typing $\pm$ α -thalassemia, Class 2: Hb E trait $\pm$ α -thalassemia, Class 3: other hemoglobinopathies).

SHAP-based interpretation

SHAP analysis of the selected model (SVM, computed with a kernel explainer) confirmed a biologically coherent decision basis. MCV was the dominant contributor, with lower values increasing the predicted probability of Class 2; MCH showed the same directional effect, followed by smaller contributions from MCHC, RDW,

NEU, and RBC. This “microcytosis with relatively preserved RBC count” signature aligns with the known Hb E phenotype and its co-inheritance with α -thalassemia and is consistent with the ReliefF ranking, supporting interpretability (Figure 4).

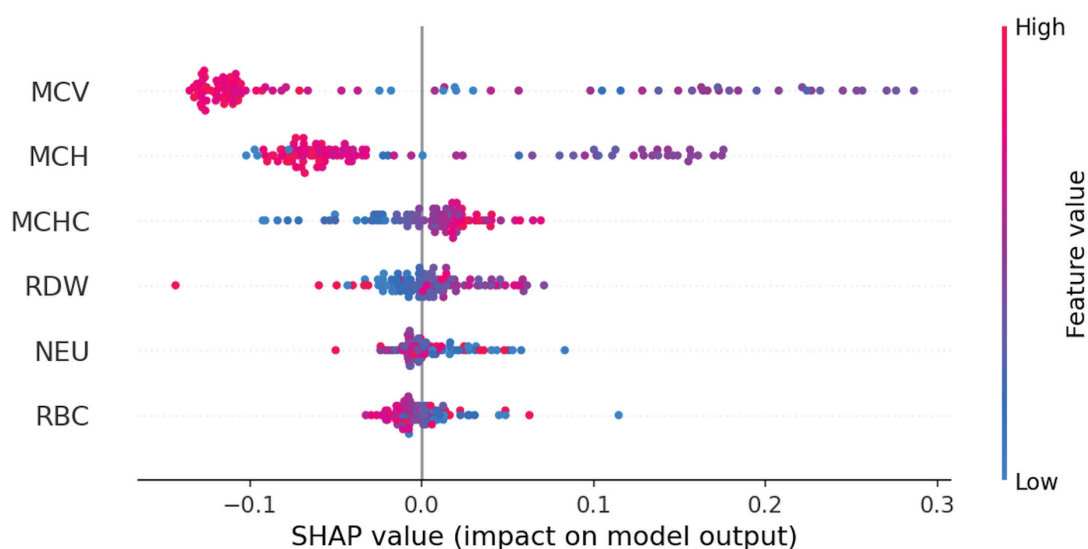


Figure 4. SHAP interpretation of the selected support vector machine (radial-basis-function kernel; SMOTE-NC pipeline) for Class 2 (Hb E trait with or without α -thalassemia), computed with a kernel explainer on internal test cases. The beeswarm plot shows the per-feature distribution of SHAP values (red: higher feature value, blue: lower). Lower MCV and MCH increase the predicted probability of Class 2 while the RBC count remains relatively preserved, consistent with the expected microcytic Hb E phenotype.

Internal test performance and the effect of SMOTE-NC

With the leakage removed, internal and external estimates for a given model were concordant, and the previously reported internal/external discrepancy disappeared: for the SVM the internal AUC (0.881) closely matched the external AUC (0.879). SMOTE-NC did not uniformly raise internal sensitivity relative to the non-SMOTE-NC pipeline (for the SVM, 70.66% with vs 70.17% without), consistent with the expected precision-coverage trade-off of oversampling. This resolves the earlier internal discrepancy, driven by synthetic samples leaking into the oversampled test set.

External validation

On the independent cohort (N=625; 181 Class 2), performance was consistent with the corrected internal estimates (Table 4). In the SMOTE-NC pipeline, the SVM achieved AUC 0.879, sensitivity 72.38%, specificity 87.39%, PPV 70.05%, NPV 88.58%, and accuracy 83.04% (Table 4); the neural network (AUC 0.886) and XGBoost (AUC 0.873) were comparable, and Naïve Bayes reached the highest default sensitivity (77.90%) at the cost of specificity. SMOTE-NC improved specificity and PPV for most models on the external set, consistent with reduced overfitting to majority-class boundaries.

Table 4. External-validation performance for Class 2 (SMOTE-NC pipeline, default operating point), all nine algorithms.

Model	AUC	Sens%	Spec%	PPV%	NPV%	Acc%
Decision Tree	0.810	51.93	88.29	64.38	81.84	77.76
Logistic Regression	0.850	65.75	86.49	66.48	86.10	80.48
Random Forest	0.867	55.80	90.99	71.63	83.47	80.80
SVM (RBF)	0.879	72.38	87.39	70.05	88.58	83.04
Gradient Boosting	0.870	62.98	89.19	70.37	85.53	81.60
XGBoost	0.873	56.91	90.99	72.03	83.82	81.12
CatBoost	0.861	61.88	89.64	70.89	85.22	81.60
Naïve Bayes	0.846	77.90	84.01	66.51	90.31	82.24
Neural Network	0.886	62.98	91.22	74.51	85.81	83.04

Sensitivity-oriented threshold tuning

Because default operating points favored specificity over sensitivity, we tuned thresholds to a $\geq 95\%$ sensitivity target on training data and applied them to the held-out sets (Table 5). For the SVM, the tuned threshold (0.080) yielded internal sensitivity 91.44% (specificity 63.85%, NPV 96.03%) and external sensitivity 91.16% (specificity

62.61%, PPV 49.85%, NPV 94.56%, accuracy 70.88%). Across models, tuned external sensitivity reached 89–93% (neural network 92.82%, Gradient Boosting 92.27%, SVM 91.16%, XGBoost 89.50%, Random Forest 88.95%), with NPV consistently $\geq 93.7\%$, appropriate for a rule-out triage tool.

Table 5. External-validation performance at the sensitivity-tuned operating point (target $\geq 95\%$ training sensitivity), SMOTE-NC pipeline.

Model	Threshold	Sens%	Spec%	PPV%	NPV%	Acc%	AUC
Neural Network	0.069	92.82	55.86	46.15	95.02	66.56	0.886
Gradient Boosting	0.080	92.27	56.76	46.52	94.74	67.04	0.870
SVM (RBF)	0.080	91.16	62.61	49.85	94.56	70.88	0.879
XGBoost	0.058	89.50	65.54	51.43	93.87	72.48	0.873
Random Forest	0.140	88.95	67.34	52.61	93.73	73.60	0.867
CatBoost	0.082	89.50	60.81	48.21	93.43	69.12	0.861
Logistic Regression	0.132	93.92	46.40	41.67	94.93	60.16	0.850
Naïve Bayes	0.005	93.37	40.99	39.21	93.81	56.16	0.846

Discussion

Hemoglobin E is among the most common structural hemoglobin variants in Thailand and Southeast Asia, where first-line screening combines red-cell indices with confirmatory HPLC or capillary electrophoresis and, when required, molecular testing.^{2,3,19} This stepwise workflow is effective but labor-intensive, and phenotypic overlap among microcytic conditions-compounded when Hb E co-occurs with α -thalassemia-can obscure Hb E trait during routine screening.^{1,19,20} We evaluated an interpretable CBC-based framework combining ReliefF feature selection, nine classifiers, SHAP interpretation, and parallel SMOTE-NC and non-SMOTE-NC pipelines, prioritizing sensitivity and NPV and adding a transparent threshold-tuning step rather than optimizing accuracy alone.

Effect of correcting the oversampling sequence

A central revision concerned the ordering of partitioning and oversampling. Applying SMOTE before splitting allows synthetic samples interpolated from training observations to enter the test set, inflating apparent performance; restricting oversampling and all preprocessing to the training partition is the recommended practice for SMOTE-family methods.^{10–12} After correction, internal discrimination was moderate-to-good (AUC ≈ 0.79 – 0.88) and concordant with external validation (reference model internal AUC 0.881 vs external 0.879), whereas the previously reported internal AUC of ~ 0.99 is best interpreted as an artifact of the earlier ordering. We therefore treat the internal-external concordance, rather than the absolute magnitude, as the informative result and refrain from claiming near-perfect discrimination.

Comparison with previous studies

Our findings are broadly consistent with, but not directly comparable to, earlier CBC-based machine-learning studies. Most prior work addressed binary problems-commonly discriminating thalassemia trait from iron-deficiency anemia-and reported high accuracy in single cohorts; ThalPred applied an SVM to red-cell indices,⁵ and later studies used ensemble and tree-based models for β -thalassemia carrier detection^{16,25} and iron-deficiency-versus-thalassemia discrimination.²¹ Studies centered on α -thalassemia have used support-vector and deep-learning approaches,^{14,22,23} and have reported feature-selection comparisons for thalassemia prediction generally.¹⁵ The present task is more demanding: a four-class problem in which the target (Hb E trait $\pm$ α -thalassemia) overlaps phenotypically with normal-typing microcytic samples, so the moderate AUCs observed are expected and should not be read as underperformance against binary benchmarks. Evaluation on unseen data, real-prevalence internal test sets, and validation in an independent external cohort guard against the over-optimistic estimates that can arise when these safeguards are absent, though cross-study differences in populations, class definitions, reference standards, and analyzers warrant caution.

Class balancing and choice of operating point

SMOTE-NC suits the mixed continuous-categorical structure of CBC data,^{13,24} and the SMOTE family is widely used to mitigate class imbalance.^{10,11} In our data, SMOTE-NC did not uniformly increase internal sensitivity-consistent with the precision-coverage

trade-off of oversampling¹²-but tended to improve external specificity and PPV, which may indicate somewhat better generalization under distribution shift. Given overlapping confidence intervals and a single external cohort, we regard this as suggestive rather than definitive. At default operating points, the models favored specificity; we therefore report a sensitivity-oriented threshold at which external sensitivity reached ~89-93% with NPV $\geq$ 93.7%, at the expected cost of specificity. Because predictive values are prevalence-dependent, local calibration and threshold selection would be needed before practical use.

Biological plausibility and interpretability

ReliefF and SHAP converged on erythrocyte indices, principally MCV and MCH, with smaller contributions from MCHC, RDW, and RBC. Lower MCV/MCH increased the predicted probability of Hb E trait while the RBC count remained relatively preserved, recapitulating the “microcytosis with maintained RBC” pattern described for Hb E and its co-inheritance with α -thalassemia.^{1,19} The concordance between statistical importance and established physiology, together with a transparent, script-based Python/Colab workflow, supports interpretability; however, SHAP values describe associations learned within this dataset and should not be interpreted as causal.

Comparison with the DCIP test

The dichlorophenol-indophenol precipitation (DCIP) test remains the principal qualitative first-line screen for Hb E in Thai laboratories.² It is a dedicated, sensitive assay for the Hb E variant, but is qualitative and read subjectively, cannot quantify Hb E or separate Hb E trait from homozygous Hb E or Hb E/ β -thalassemia, does not detect α -thalassemia, and requires dedicated reagents and manual processing. The CBC-based model is complementary rather than competing; it reuses data already generated for every patient at no additional cost and is fully reproducible, but at default thresholds its sensitivity is lower than that expected of DCIP. We therefore do not propose it as a replacement for DCIP or hemoglobin typing, but as an automated, low-cost pre-screen that prioritizes samples for confirmatory testing.

Distinguishing Hb E trait with versus without α -thalassemia

Separating Hb E trait with α -thalassemia from Hb E trait alone using CBC data would be clinically valuable, since hemoglobin typing cannot reliably resolve this and molecular testing is often inaccessible and costly.²⁰ Both subgroups are microcytic and overlap substantially, with co-inherited α -thalassemia further lowering MCV, MCH, and the Hb E fraction,^{1,20} so their separation from indices alone is intrinsically limited. In this cohort, we combined the subgroups within Class 2 because molecular α -genotyping was unavailable for enough cases to train a reliable sub-classifier; we

present the merged target rather than making claims our data cannot support. With molecularly confirmed labels, a hierarchical model could be evaluated in the future, potentially aided by genotype- or SNP-informed features.^{18,23}

Distinguishing Class 0 from Class 1

Class 0 and Class 1 share a normal typing pattern and are indistinguishable by typing alone. Class 1 additionally includes silent or α -thalassemia carriers who may be microcytic despite normal typing, whereas Class 0 comprises normal or non-clinically significant samples.²⁰ Because α -thalassemia produces no abnormal band, the distinction rests on red-cell indices and, definitively, on molecular testing. This overlap likely accounts for the residual misclassification of Class 2 into the adjacent normal-typing groups seen in the confusion matrices, an expected consequence of shared microcytic phenotypes.

Limitations

This study is retrospective and single-center and relies on CBC-derived features without iron studies, which are informative when distinguishing thalassemia from iron deficiency.^{7,8} Labeling was heterogeneous across adjacent normal-typing classes, Class 2 combined Hb E trait with and without α -thalassemia, and predictive values are prevalence-dependent, limiting transportability of PPV and NPV. Although external validation is a strength, a single external cohort does not establish multi-site generalizability, and the modest default sensitivity means the model is best regarded as a triage aid rather than a stand-alone screen.

Future study

Future work should pursue multi-center validation, analyzer harmonization, integration of iron indices and hemoglobin-fraction data, molecularly labeled datasets for α -status sub-classification, and prospective, prevalence-aware calibration before deployment.

Conclusion

Using a corrected, leakage-free pipelin-splitting the data before applying SMOTE-NC and fitting all preprocessing on training data only-an explainable model based on routine CBC parameters provides realistic and reproducible performance for identifying the Hb E trait. Among nine algorithms, a support vector machine gave the best combination of sensitivity and discrimination and was selected as the primary model. Internal and external discrimination are now concordant (AUC $\approx$ 0.88 for the selected model), and SHAP confirms a biologically coherent, MCV/MCH-driven decision basis. At a sensitivity-oriented operating point, the model reaches approximately 90% sensitivity with high negative predictive value on an independent cohort, supporting its use as a low-cost, objective pre-screening and triage layer that prioritizes samples

for DCIP testing and hemoglobin typing rather than replacing them. With local calibration and, in future, molecularly labeled data to resolve α -thalassemia co-inheritance, this approach can help streamline thalassemia screening workflows in high-prevalence, resource-limited settings.

Ethical approval

The Human Research Ethics Committee of Thammasat University (Science), Thailand (HREC-TUSc), reviewed and approved this study (Certificate of Approval No. 120/2567, Project No. 67AH137). Ethical approval was granted on November 7, 2024, with approval validity through November 6, 2025 (extended to November 6, 2026).

This study was reviewed and approved by the Institutional Review Board of Bamrasnaradura Infectious Diseases Institute, Certificate of Approval No. IRB/BIDI S012b/67_ExpD. Ethical approval was granted on May 9, 2024, with approval validity through May 8, 2025 (extended to May 8, 2026).

Declaration of generative AI and AI-assisted technologies

During the preparation of this article, the authors used ChatGPT to proofread and optimize parts of the text. The authors then reviewed and edited the content as needed, followed by final proofreading by a native English speaker. The authors take full responsibility for the content of this publication.

Funding

This study was supported by the Thammasat Research Unit in Medical Technology and Precision Medicine Innovation, Faculty of Allied Health Sciences, Thammasat University.

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.

CRedit authorship contribution statement

Chollanot Kaset: writing-original draft, methodology, conceptualization, software, investigation; **Tanita Rerkchalerm:** writing-review and editing, validation; **Tanatorn Tanantong:** validation, writing-review and editing, supervision; **Sirinart Chomean:** project administration, validation, writing-original draft, software, investigation.

Acknowledgements

This research is supported by Thailand Science Research and Innovation (TSRI) Fundamental Fund, Fiscal year 2025. This research gratefully acknowledges the Medical Technology and Reference Laboratory for Infectious Disease (MTRL), Bamrasnaradura Infectious Diseases Institute, for providing the valuable data.

References

- [1] Pornprasert S, Tookjai M, Punyamung M, Pongpunyayuen P. HbE level and red cell parameters in heterozygous HbE with and without -thalassemia trait. *Indian J Hematol Blood Transfus.* 2018; 34(4): 662-5. doi.org/10.1007/s12288-018-0947-8.
- [2] Paiboonsukwong K, Jopang Y, Winichagoon P, Fucharoen S. Thalassemia in Thailand. *Hemoglobin.* 2022; 46(1): 53-7. doi.org/10.1080/03630269.2022.2025824.
- [3] Nuntakarn L, Fucharoen S, Fucharoen G, Sanchaisuriya K, Jetsrisuparb A, Wiangnon S. Molecular, hematological and clinical aspects of thalassemia major and thalassemia intermedia associated with Hb E--thalassemia in Northeast Thailand. *Blood Cells Mol Dis.* 2009; 42(1): 32-5. doi.org/10.1016/j.bcmd.2008.09.002.
- [4] Yamsri S, Singha K, Prajantasen T, Fucharoen G, Sanchaisuriya K, Fucharoen S. A large cohort of -thalassemia in Thailand: molecular, hematological and diagnostic considerations. *Blood Cells Mol Dis.* 2015; 54(2): 164-9. doi.org/10.1016/j.bcmd.2014.11.008.
- [5] Laengsri V, Shoombuatong W, Adirojananon W, Nantasenamart C, Prachayasittikul V, Nuchnoi P. ThalPred: a web-based prediction tool for discriminating thalassemia trait and iron deficiency anemia. *BMC Med Inform Decis Mak.* 2019; 19(1): 212. doi.org/10.1186/s12911-019-0929-2.
- [6] Appiahene P, Asare JW, Donkoh ET, Dimauro G, Maglietta R. Detection of iron deficiency anemia by medical images: a comparative study of machine learning algorithms. *BioData Min.* 2023; 16(1): 18. doi.org/10.1186/s13040-023-00319-z.
- [7] Plengsuree S, Punyamung M, Yanola J, Suwankesawong W, Pornprasert S. Red cell indices and formulas used in differentiation of -thalassemia trait from iron deficiency in Thai adults. *Hemoglobin.* 2015; 39(4): 235-9. doi.org/10.3109/03630269.2015.1048352.
- [8] Jahangiri M, Rahim F, Malehi AS. Diagnostic performance of hematological discrimination indices to discriminate between -thalassemia trait and iron deficiency anemia and using cluster analysis: introducing two new indices tested in an Iranian population. *Sci Rep.* 2019; 9(1): 18607. doi.org/10.1038/s41598-019-54575-3.
- [9] Ayyildiz H, Arslan Tuncer S. 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. *Chemom Intell Lab Syst.* 2020; 196: 103886. doi.org/10.1016/j.chemolab.2019.103886.
- [10] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res.* 2002; 16: 321-57. doi.org/10.1613/jair.953.

- [11] Pradipta GA, Wardoyo R, Musdholifah A, Sanjaya INH, Ismail M. SMOTE for handling imbalanced data problem: a review. In: Proceedings of the 6th International Conference on Information and Computing (ICIC); 2021 Nov 18-19; Semarang, Indonesia. Piscataway (NJ): IEEE; 2021. p. 1-6. doi.org/10.1109/ICIC54025.2021.9632912.
- [12] Elreedy D, Atiya AF, Kamalov F. A theoretical distribution analysis of synthetic minority over-sampling technique (SMOTE) for imbalanced learning. Mach Learn. 2024; 113(7): 4903-23. doi.org/10.1007/s10994-022-06296-4.
- [13] Islahulhaq I, Wibowo W, Ratih ID. Classification of non-performing financing using logistic regression and SMOTE-NC. Int J Adv Soft Comput Appl. 2021; 13(3): 115-28. doi.org/10.15849/ijasca.211128.09.
- [14] Nasir MU, Naseem MT, Ghazal TM, Akram MU, Ashraf I. A comprehensive case study of deep learning on the detection of alpha and beta thalassemia using public and private datasets. Sci Rep. 2025; 15(1): 8540. doi.org/10.1038/s41598-025-97353-0.
- [15] Saleem M, Aslam W, Lali MIU, Rauf HT, Nasr EA. Predicting thalassemia using feature selection techniques: a comparative analysis. Diagnostics (Basel). 2023; 13(22): 3441. doi.org/10.3390/diagnostics13223441.
- [16] Das R, Saleh S, Nielsen I, Sadiq R, Datta C, Chakrabarti P, et al. Performance analysis of machine learning algorithms and screening formulae for -thalassemia trait screening of Indian antenatal women. Int J Med Inform. 2022; 167: 104866. doi.org/10.1016/j.ijmedinf.2022.104866.
- [17] Long Y, Bai W. Constructing a novel clinical indicator model to predict the occurrence of thalassemia in pregnancy through machine learning algorithm. Front Hematol. 2024; 3: 1341225. doi.org/10.3389/frhem.2024.1341225.
- [18] Munkongdee T, Tongsima S, Ngamphiw C, Sonsilphong S, Tanjarern S, Winichagoon P, et al. Predictive SNPs for -thalassemia/HbE disease severity. Sci Rep. 2021; 11(1): 10229. doi.org/10.1038/s41598-021-89641-2.
- [19] Fucharoen S, Weatherall DJ. The hemoglobin E thalassemias. Cold Spring Harb Perspect Med. 2012; 2(8): a011734. doi.org/10.1101/cshperspect.a011734.
- [20] Polprasert C, Wongprachar P, Suksusut A, Kanjanabuch T, Vaneesorn J, Tanakanya S, et al. Comprehensive screening for coexisting heterozygous -thalassemia in hemoglobin E trait. Hematology. 2020; 25(1): 276-9. doi.org/10.1080/16078454.2020.1786972.
- [21] Tepakhan W, Srisintorn W, Penglong T, Saelue P. Machine learning approach for differentiating iron deficiency anemia and thalassemia using random forest and gradient boosting algorithms. Sci Rep. 2025; 15(1): 4585. doi.org/10.1038/s41598-025-01458-5.
- [22] Lachover-Roth I, Peretz S, Zoabi H, Shani-Gigi N, Abu-Kaf H, Koren A, et al. Support vector machine-based formula for detecting suspected -thalassemia carriers. Int J Mol Sci. 2024; 25(12): 6446. doi.org/10.3390/ijms25126446.
- [23] Christensen F, Kılıç DK, Nielsen IE, Vinding MS, Farooq F, Winther O, et al. Classification of -thalassemia data using machine learning models. Comput Methods Programs Biomed. 2025; 260: 108581. doi.org/10.1016/j.cmpb.2024.108581.
- [24] Mukherjee M, Khushi M. SMOTE-ENC: a novel SMOTE-based method to generate synthetic data for nominal and continuous features. Appl Syst Innov. 2021; 4(1): 18. doi.org/10.3390/asi4010018.
- [25] Rustam F, Ashraf I, Jabbar S, Ullah S, Sultana R, Jamil F, et al. Prediction of -thalassemia carriers using complete blood count features. Sci Rep. 2022; 12(1): 17616. doi.org/10.1038/s41598-022-22011-8.