Early Version – Uncorrected



Developing a Multimodal Deep Learning Framework for Classifying Thalassemia and Iron Deficiency Anemia: A Conceptual Step towards Digital-Assisted Decision-Making in Hematology

Warit Ruanglertboon, Chutima Cheranakhorn, Komsai Suwanno, Satanun Sottisataworn, Yada Suwannachote, Marisa Meethong, Sorawat Sangkeaw, Preecha Vonghirandecha

DOI: http://dx.doi.org/10.31584/jhsmr.20261514

Received: 12 October 2025

Revised: 17 March 2026

Accepted: 23 May 2026

Published online: 16 November 2026

Please cite this article as: Ruanglertboon W, Cheranakhorn C, Suwanno K, Sottisataworn S, Suwannachote Y, Meethong M, et al. Developing a multimodal deep learning framework for classifying thalassemia and iron deficiency anemia: a conceptual step towards digital-assisted decision-making in hematology. J Health Sci Med Res 2026 doi: http://dx.doi.org/10.31584/jhsmr.20261514

Early Version — Uncorrected: This accepted manuscript has not been typeset or proofread. Content and formatting may change prior to final publication and should not be regarded as the final Version of Record.

© 2026 JHSMR. Hosted by Prince of Songkla University.

This is an open access article under the CC BY-NC-ND license

(https://creativecommons.org/licenses/by-nc-nd/4.0/)

Original Article

Developing a Multimodal Deep Learning Framework for Classifying Thalassemia and Iron Deficiency Anemia: A Conceptual Step towards Digital-Assisted Decision-Making in Hematology

Warit Ruanglertboon[1,2], Chutima Cheranakhorn, M.D.[3], Komsai Suwanno, M.D.[3], Satanun Sottisataworn, M.D.[3], Yada Suwannachote[4], Marisa Meethong[4], Sorawat Sangkeaw, M.D.[5], Preecha Vonghirandecha[4]

[1]Research Center in Mathematics and Statistics with Applications, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla 90110, Thailand.
[2]Pharmacology Program, Division of Health and Applied Science, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla 90110, Thailand.
[3]Department of Internal Medicine, Hatyai Hospital, Hat Yai, Songkhla 90110, Thailand.
[4]Division of Computational Science, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla 90110, Thailand.
[5]Department of Social Medicine, Hatyai Hospital, Hat Yai, Songkhla 90110, Thailand.

***Corresponding Author:**
Preecha Vonghirandecha
Division of Computational Science, Faculty of Science, Prince of Songkla University, Songkhla 90110, Thailand.
E-mail: preecha.v@psu.ac.th

Early Version - Uncorrected

Abstract:

Objectives: Anemia is one of the most common hematological disorders and is recognized by the WHO as a major global health concern. Differentiating between anemia subtypes, such as thalassemia and iron deficiency anemia, is crucial for guiding the appropriate treatment. This study aimed to develop a multimodal deep learning framework to support digital-assisted decision-making in hematological disorders.

Material and Methods: Patient characteristics and corresponding peripheral blood smear (PBS) images were retrieved from electronic medical records under ethical approval. The proposed framework consisted of two components. The first component comprised three predictive layers: image analysis, patient characteristics, and patient profile similarity. The second component integrated these predictions through a majority voting mechanism, whereby concordance of two out of three layers determined the final classification.

Results: A total of 166 patients and 764 high-resolution PBS images were included, comprising 425 from thalassemia and 339 from iron deficiency anemia. Within the first component, EfficientNetB0 yielded the best performance for image analysis, the decision tree performed best for patient characteristics, and k-nearest neighbor provided optimal results for profile similarity. Combining these three models produced the most reliable predictive performance for the dataset.

Discussion: The multimodal approach achieved superior performance compared with individual or dual algorithms. Nevertheless, larger datasets are required to further enhance prediction accuracy and support clinical translation.

Conclusion: The multimodal deep learning framework developed in this study demonstrated promising predictive performance. Further refinement under clinical supervision may advance the integration of image analysis into hematology practice.

Keywords: Deep learning, thalassemia, anemia, decision-making

Early Version - Uncorrected

Introduction

Anemia is a prevalent hematological disorder characterized by decreased levels of red blood cells (RBCs) or hemoglobin (Hb). According to the World Health Organization (WHO), anemia remains a major global health concern, particularly affecting women and young children. In 2023, the global prevalence of anemia among women aged 15–49 years was reported at 30.7%, with a higher prevalence observed in pregnant women (35.5%) compared to non-pregnant women (30.5%). In Thailand, WHO data from 2023 indicated that 20.7% of women aged 15–49 years were affected by anemia, with the prevalence rising to 23.8% among pregnant women in the same age group. Anemia is conventionally classified into three morphological types based on red cell indices: hypochromic microcytic, normochromic normocytic, and macrocytic anemia. Among these, thalassemia and iron deficiency anemia (IDA) are the leading causes of hypochromic microcytic anemia, especially in Southeast Asia. In central Thailand, IDA is the primary cause of anemia in young women, while thalassemia, a common genetic disorder, significantly contributes to the region's anemia burden [1,2]. The prevalence of IDA among Thai women of reproductive age was reported as 28.7% in 2013, increasing to 30.2% in 2014 and 31.8% in 2015 [1]. Furthermore, a separate study found that 21% of educated young Thai women were anemic, with thalassemia (28%) and IDA (21%) identified as the principal etiologies [2, 3]. Misclassification between thalassemia and IDA can lead to inappropriate treatment. Administering iron to thalassemia carriers may cause iron overload, while undiagnosed IDA may delay detection of underlying conditions such as gastrointestinal bleeding [4, 5, 6].

In parallel, machine learning and deep learning have thrived in recent years, driven by advances in computational power and accessibility. The core mechanism of these approaches, training mathematical algorithms to predict unseen data during testing, forms the foundation of numerous modern applications across diverse domains. A particular area of interest is image analysis, where pixel-level data are extracted to identify patterns for subsequent applications. Importantly, there have been numerous attempts to integrate deep learning into practical diagnostic applications, particularly in fields such as radiology, pathology, dermatology, and ophthalmology [7, 8, 9, 10]. Essentially, the predictive performance of deep learning has been shown to achieve outcomes comparable to those of human experts. In cytological image analysis, techniques such as combined edge-selection across individual RGB channels have been shown to enhance the

Early Version - Uncorrected

extraction of morphological features from cellular images [11, 12]. This highlights the potential role of machine-assisted clinical diagnosis in contemporary medicine.

Within the machine learning paradigm, classification techniques are employed to identify patterns and determine whether an input belongs to a specific class. Some algorithms, such as k-nearest neighbors (KNN), classify inputs based on distances between data points, while others, such as support vector machines (SVM), construct an optimal decision boundary by maximizing the margin between classes. Conversely, rule-based classification methods, such as decision trees, operate under predefined if–then rules to derive outcomes. In contrast, statistical approaches such as linear discriminant analysis (LDA) classify inputs based on probabilistic models and the linear separation between classes. In the realm of deep learning, convolutional neural network (CNN) architectures are widely used for image analysis tasks. Early models such as AlexNet introduced stacked convolutional and fully connected layers, establishing the foundation of modern CNNs. ResNet, through the use of residual connections, enabled the training of much deeper networks by addressing the vanishing gradient problem. MobileNetV2, in contrast, was designed as a lightweight model for resource-constrained environments, employing depth-wise separable convolutions and inverted residual blocks to reduce computational cost. EfficientNetB0 serves as the baseline model of the EfficientNet family, which employs a compound scaling method to uniformly balance network depth, width, and resolution across different model variants, thereby achieving a superior tradeoff between accuracy and efficiency compared to earlier CNN architectures.

Against this background, the present study aimed to develop a unified framework for hematological prediction by integrating machine learning with deep learning approaches. Patient characteristics and laboratory data were modeled using machine learning algorithms, while corresponding image data were analyzed with deep learning models. In addition, similarity retrieval based on patient records was incorporated as a complementary component prior to consolidating outcomes across parameters. The overarching goal was to derive a robust predictive framework for distinguishing thalassemia from IDA, thereby supporting clinical decision-making in hematological diagnosis.

Early Version - Uncorrected

Material and Methods

Demographic data

The current study utilized peripheral blood smear (PBS) images obtained from a data repository under an agreement with Hatyai Hospital. Data from patients who met the specified criteria were selected for inclusion. The inclusion criteria were as follows:

1. Age 16 years or older
2. Diagnosed with either thalassaemia or iron deficiency
3. Availability of a high-resolution PBS image (1944 × 2592 pixels)

The exclusion criteria included individuals with anemia due to other underlying conditions, such as aplastic anemia, chronic anemia, autoimmune hemolytic anemia, multiple myeloma, and leukemia. Additionally, patients diagnosed with both thalassemia and IDA concurrently, as well as those with stage 3 or higher chronic kidney disease, were excluded to minimize confounding factors and preserve the validity of the study. Available haematological parameters, as well as baseline characteristics, were collected according to hospital ID. This study was approved by the institutional ethics committee of Hatyai Hospital (HYH EC 002-67-01). The study utilized retrospective clinical data retrieved from the hospital electronic medical record system. All patient data were de-identified prior to analysis by removing identifiable information and replacing it with study-specific identifiers.

PBS images

PBS samples were prepared and stained using Wright stain according to routine laboratory procedures. Smear interpretation was initially performed by a clinical pathology resident and subsequently reviewed and approved by a supervising hematologist before the final laboratory report was issued. The ground-truth diagnostic labels used in this study were derived from these confirmed laboratory diagnoses recorded in the hospital electronic medical record system. PBS images were captured during the same laboratory examination using a Leica ICC50 W microscope camera under the supervision of a clinical pathologist. Only high-resolution color PBS images were included for analysis. The time interval between image capture and confirmation by the hematologist ranged from 30–40 minutes to no longer than 48 hours, depending on the workload.

Early Version - Uncorrected

Construction of multimodal deep learning framework

The overall architecture of the proposed multimodal deep learning framework is presented in Figure 1. The framework was designed with two compartments: the first consisting of three layers, and the second serving as the pre-prediction decision stage.

The first layer employed image-based analysis using deep convolutional neural networks (DCNNs). Four architectures were evaluated: AlexNet, MobileNetV2, ResNet50, and EfficientNetB0. Model performance was assessed using partitioned datasets, with a ratio of 60:20:20 for training, validation, and testing, respectively.

The second layer incorporated baseline characteristics and haematological profiles as numerical and categorical predictors. Four conventional machine learning models, including SVM, decision tree, Naïve Bayes (NB), and LDA, were applied to evaluate their predictive performance.

The third layer utilized a classification algorithm. The k-nearest neighbor (KNN) method was applied to compare the similarity of new patients with historical records. Relevant baseline characteristics and hematological profiles were normalized to the range [0,1] prior to similarity computation. The value of k was chosen to provide the most balanced trade-off among patient records, yielding the highest similarity. For each new patient, the algorithm retrieved the most similar historical records of k and assigned the output class label accordingly. To evaluate the robustness of the similarity-based KNN layer, several alternative distance metrics (Euclidean, Manhattan, Cosine, and Mahalanobis) were assessed. The metric yielding the highest classification accuracy was selected for the final framework.

The second compartment of the framework integrates the outputs from the three preceding layers. Two ensemble schemes were evaluated: (i) pairwise combinations of layers (nCr (3, 2) = 3 combination models), in which predictions were derived from each pair, and (ii) majority voting, the fusion-based majority voting (FBMV) across all three layers. In FBMV, each classifier contributed one vote, and a sample was assigned to the class label that received at least two of the three votes. For pairwise model fusion, a performance-based weighted voting scheme was applied to account for differences in predictive strength between modalities. Model weights in the ensemble were derived through performance normalization. Specifically, each model weight was calculated by dividing its evaluation metric (accuracy or F1-score) by the sum of the corresponding metrics of the models included in the ensemble. Because both metrics produced very similar weight distributions across the modalities, the accuracy-based weights were adopted for the final

Early Version - Uncorrected

pairwise fusion for simplicity and consistency. In contrast, the final three-modality integration employed a majority voting mechanism, where agreement between any two of the three modalities determined the final classification.

Performance evaluation

All models across all layers were evaluated and compared based on their predictive performance. The assessment included accuracy, precision, recall, F1-score, and loss. Accuracy represents the overall proportion of correctly classified samples. Precision measures how many of the positive predictions were truly positive, while recall reflects how many of the actual positive cases were correctly identified. The F1-score combines both precision and recall into a single metric, providing a balanced view of model performance. Finally, the loss quantifies the prediction error that arises during training, where lower values indicate better optimization.

Predictive performance was assessed using MATLAB (version R2024b), while all other statistical analyses were conducted in R (version 4.5.1) with the required computational packages.

Supplementary analysis

A web-based application was deployed as part of the preliminary evaluation. The application was planned for local deployment to collect feedback from clinicians for further refinement. Details of technical implementation are provided in the Supplementary Technical Information.

Results

The demographic data following validation of the inclusion criteria are presented in Table 1. A total of 166 patients were included, with 425 PBS images from thalassemia patients and 339 images from those with IDA available for analysis.

Deep learning model performance

The performance of the evaluated deep learning models is summarized in Figure 2 and Supplementary Figures S1 and S2. Among the four models, EfficientNetB0 demonstrated the best overall performance across most evaluation metrics, recording the highest accuracy (0.955),

Early Version - Uncorrected

precision (0.966), and F1-score (0.949). It also yielded the lowest loss (0.131), indicating the smallest optimization error during training. In comparison, AlexNet exhibited the lowest performance in terms of F1-score. Although MobileNetV2 achieved the highest recall, its overall performance was slightly lower across several other evaluation metrics compared with EfficientNetB0. ResNet50 performed relatively well but still fell short of EfficientNetB0 across the evaluated indicators. These findings highlight the superior capability of EfficientNetB0 in classifying PBS images within the proposed framework; therefore, it was selected for incorporation into the actual multimodal learning framework. Model calibration was further assessed for the best-performing model (EfficientNetB0). The model achieved a Brier score of 0.0459, indicating good probabilistic calibration. The reliability curve (Supplementary Figure S3) further demonstrated that the predicted probabilities were generally aligned with the observed event frequencies across probability bins.

To further ensure that model evaluation was not influenced by potential image-level overlap, additional validation using patient-wise data separation was conducted. Receiver operating characteristic (ROC) curves and precision–recall (PR) curves for the evaluated models are presented in Supplementary Figures S4–S5 and Supplementary Table S2, together with the corresponding 95% CI. These analyses confirmed that EfficientNetB0 maintained the strongest discrimination performance among the evaluated models under patient-level validation. To explore potential subgroup bias, model performance was evaluated across demographic strata including sex and age groups. The detailed results of the subgroup analysis are presented in Supplementary Table S1.

Figure 2. Performance of the four deep learning models.

Classification model performance

Figure 3 summarizes the performance of the evaluated classification models. The decision tree outperformed LDA and NB across all metrics, achieving the highest accuracy (0.933), F1-score (0.931), precision (0.959), and recall (0.904). It also yielded the lowest loss (0.067), confirming its superior predictive capability among the three models. Although SVM performed comparably to the decision tree in recall, its overall performance was the lowest. Therefore, the

Early Version - Uncorrected

decision tree was selected for incorporation into the subsequent multimodal deep learning framework.

Optimization of k and similarity performance

The performance of the KNN algorithm across different values of *k* is shown in **Figure 4**. Overall, thalassemia cases were classified with consistently higher accuracy compared with IDA. Accuracy for thalassemia ranged from 85.1% (*k* = 1, 2) to 92.6% (*k* = 7, 9), with misclassification rates between 7.4% and 14.9%. Meanwhile, accuracy for IDA was lower, ranging from 44.2 % (*k* = 13, 14) to 65.4% (*k* = 1, 2), with higher misclassification rates (up to 55.8%). The most balanced trade-off was observed at *k* = 8, where classification accuracy reached 91.5% for thalassemia and 61.5% for IDA, with corresponding misclassification rates of 8.5% and 38.5%, respectively.

Combined model performance

The performance of the combined models is presented in Figure 5. In total, three pairwise combined models and one FBMV model were compared, along with the best-performing model selected from the prior evaluation. Herein, FBMV achieved the highest accuracy (0.966), F1-score (0.967), and recall (0.989), while also recording the lowest loss (0.034). By contrast, the decision tree yielded the highest precision (0.959), indicating its relative strength in minimizing false positives. These results highlight FBMV as the most robust model overall, with the decision tree contributing complementary precision performance.

The optimized multimodal deep learning framework

Figure 6. An illustration of the multimodal deep learning framework.

Following the selection of the best-performing models from the three layers, the final multimodal deep learning framework is presented in Figure 6. This framework was implemented and deployed locally for physicians to test and provide feedback, supporting ongoing development to enhance reliability and expand functionality. A prototype of the web application is shown in Supplementary Technical Information.

Discussion

The current era is enriched with a wide range of technologies, many of which have the potential to enhance medical frameworks. This study demonstrates how such advancements can

Early Version - Uncorrected

be pragmatically incorporated to support clinical decision-making. Traditional practice, which relies heavily on human evaluation and diagnostic experience through image interpretation, can now be augmented by digital assistance in the form of deep learning mechanisms.

In the context of PBS images, the majority of red blood cells in both thalassemia and IDA cases exhibit similar morphological characteristics, making visual differentiation inherently challenging. To address this issue, a multimodal deep learning framework was constructed to alleviate these difficulties. The proposed framework consists of three layers, with its uniqueness lying in the deliberate and systematic model construction. Rather than relying on arbitrary model selection, each algorithm was chosen to align with the specific attributes of the target data. Designed to classify images as either thalassemia or IDA, the framework begins with a first layer oriented toward image input, for which deep learning models represent the most appropriate paradigm. Based on 5-fold cross-validation, EfficientNetB0 exhibited the highest performance across all metrics and was therefore selected as the first layer to extract information from image-based input data. In addition to predictive accuracy, calibration analysis demonstrated that EfficientNetB0 produced well-calibrated probability estimates, with a Brier score of 0.0459 and a reliability curve showing reasonable agreement between predicted and observed outcomes. Such calibration is important for potential clinical decision-support applications where probability estimates may guide diagnostic interpretation. The performance observed in this study is consistent with the previous literature on image analysis tasks. EfficientNet belongs to the convolutional neural network family, and the B0 variant is notable for having fewer parameters than other variants such as the B7, which alleviates the risk of overfitting. Comparative research on lung abnormality imaging has likewise demonstrated the superiority of EfficientNetB0 over ResNet50 [13]. Similar findings have been reported in other domains, where EfficientNetB0 has demonstrated its capability in image classification tasks such as diabetic retinopathy detection, colonoscopy-based diagnosis of precancerous cervical lesions, and even biometric applications, for example, fingerprint recognition [8, 14, 15]. Furthermore, additional validation using patient-wise data separation confirmed that predictive performance remained consistent (Supplementary Table S2), suggesting that it was not driven by image-level overlap and supporting the robustness of the proposed framework. The reduction in model performance when shifting from image-based to patient-wise data separation is expected, as similar patterns have been reported in previous studies

Early Version - Uncorrected

[16]. However, this decrease more accurately reflects the model's real-world predictive capability, avoiding overly optimistic estimates and leaving room for further refinement and improvement.

Subgroup analysis suggested some variation in predictive performance across demographic groups. For example, model accuracy was slightly higher in males compared with females. Such variation may reflect differences in the sample distribution or underlying clinical heterogeneity within the dataset. The pediatric subgroup was not included in the stratified analysis due to the limited number of cases, which could lead to unstable estimates. Future studies involving larger, more balanced datasets will be important to further assess any potential subgroup bias, ensuring fairness across clinical populations.

While the first layer focused on image input, the second layer utilized clinical parameters and baseline characteristics from the corresponding patients as input for prediction. Four classification models (SVM, LDA, NB, and decision tree) were evaluated at this stage, as these algorithms are widely applied in clinical diagnostic tasks [17, 18]. Among these, the decision tree demonstrated the best predictive performance, achieving the highest scores across all metrics with the lowest loss. This outcome may be attributed to the greater flexibility of tree-based models in capturing complex feature relationships [19].

To further examine the potential risk of overfitting in the decision tree model, additional experiments were conducted using several pruning levels. As pruning increased, overall classification performance gradually declined. The unpruned tree achieved the highest accuracy, recall, precision, and F1-score among the tested configurations, whereas progressively pruned trees showed reduced predictive performance (Supplementary Table S3). These findings suggest that additional pruning did not improve predictive performance for the current dataset. Therefore, the unpruned decision tree was retained as the tabular component of the proposed multimodal framework. Nevertheless, tree pruning strategies and ensemble methods, such as random forests or gradient boosting, particularly when trained on an expanded cohort, may further improve robustness and reduce the risk of overfitting. These approaches will be explored in future studies with additional patient recruitment and larger sample sizes. By contrast, the relatively lower performance of LDA may stem from its assumption of linear class boundaries and equal covariance across groups, which may not hold in general real-world clinical data [20]. Similarly, NB is limited by its independence assumption [21], which is often unrealistic for clinical variables that are highly correlated, thereby reducing predictive accuracy.

Early Version - Uncorrected

The third layer employed similarity retrieval using the KNN algorithm. Patient profiles were compared with records in the clinical database to identify the closest matches. Once the most similar cases were identified, the corresponding diagnoses of those records were retrieved and provided as the predicted outcome. Similarity was determined by calculating the Euclidean distance between patient profiles, and the final prediction was generated based on the weighted combinations of the nearest neighbors. The application of this approach has been reported in previous literature [22], supporting the appropriateness of incorporating KNN into the present framework. In this study, $k = 8$ was selected following experimental optimization. The choice of k reflected a balance between classification accuracy and misclassification rates, as improvements in one domain sometimes came at the expense of performance in another. While certain values of k showed higher accuracy for individual classes, $k = 8$ provided the most balanced trade-off, yielding the highest overall accuracy with reduced error rates across both thalassemia and iron deficiency anemia.

To further evaluate the robustness of the similarity-based KNN layer, a sensitivity analysis was conducted by comparing several distance metrics, including Euclidean, Manhattan, Cosine, and Mahalanobis distances (Supplementary Table S4). Among the evaluated metrics, Euclidean distance achieved the highest classification accuracy (0.8356), followed by Cosine (0.8288), Manhattan (0.8219), and Mahalanobis distance (0.8219). These findings support the use of Euclidean distance for the current dataset. Nevertheless, the relatively small differences in performance across the evaluated metrics suggest that the KNN component remained reasonably stable across alternative distance formulations. Future studies with larger datasets may further explore clinically informed feature weighting or learned distance functions to improve robustness, particularly for iron deficiency anemia classification.

Following the evaluation of all three layers, the final step in the proposed framework was the prediction calculation. The framework consisted of two compartments: the first served as the backbone of the prediction process, while the second consolidated the outputs to derive the final outcome. Prediction performance against the ground truth was assessed by first evaluating each layer individually, then examining pairwise combinations of two layers, and finally testing the combined performance of all three layers. The results showed that utilizing all three layers, with the prediction outcome determined when at least two layers were in agreement, provided the most reliable performance. Such practices have been increasingly adopted in recent studies, showing

Early Version - Uncorrected

improved performance compared to the use of individual deep learning algorithms alone [8, 23, 24]. In the pairwise fusion setting, a performance-based weighted voting scheme was applied to account for the differences in predictive strength between the two modalities. This approach is particularly relevant when combining two classifiers, where equal voting may allow a weaker model to exert the same influence as a stronger one. By assigning weights proportional to model performance, the decision process better reflects the relative reliability of each modality. In contrast, the final three-layer integration relied on majority voting, where agreement between two of the three classifiers determines the final prediction and reduces the influence of a single weaker modality. Similarly, the proposed multimodal framework may offer advantages over traditional index-based discrimination methods used for differentiating thalassemia and iron deficiency anemia, particularly in reducing misclassification of IDA cases [25].

A major strength of this study was the use of real-world data, incorporating both patient information and PBS images. Validation processes were conducted in collaboration with clinical specialists to ensure data integrity. In addition, the availability of a sufficiently large image dataset provided robust support for the reliability of the results. Moreover, this research provides a glimpse of practical implementation through the development of a web-based application that incorporates the proposed framework as the core prediction engine, as shown in Supplementary Figures S6 to S9. The scope of comparison in this study was limited by the relatively small number of machine learning and deep learning algorithms evaluated. The present study used data from a single institutional repository; therefore, external or temporal validation was not performed. Future studies incorporating multi-center datasets will be important to further evaluate the generalizability of the proposed framework.

The proposed framework is intended to function as a clinical decision-support tool to assist physicians and laboratory personnel during the interpretation of PBS images rather than replacing definitive diagnostic procedures. In clinical practice, misclassification between thalassemia and iron deficiency anemia may lead to inappropriate treatment decisions. For example, false-positive predictions for thalassemia may prompt unnecessary follow-up investigations, while false-negative cases could delay appropriate treatment for IDA. By incorporating a computational tool as an initial screening support, the framework may help reduce cognitive workload for laboratory personnel and physicians by highlighting cases that require closer examination before formal interpretation of PBS findings. However, to improve the feasibility of implementing such tools in

Early Version - Uncorrected

clinical practice, further studies involving larger sample sizes, more diverse patient populations stratified by baseline characteristics, and prospective real-world evaluation alongside laboratory personnel will be necessary.

Conclusion

The proposed multimodal deep learning framework integrates EfficientNetB0 for image data, a decision tree for demographic and laboratory data, and KNN with $k = 8$ for patient similarity retrieval. Combined through a majority voting mechanism, this approach demonstrated high predictive performance for PBS image classification.

References

1. WHO. Trends in anemia in women and children: 1995 to 2016; WHO 20172017. Available from: http://www.WHO.int/data/ gho/data/indicators/indicator-details/GHO/prevalence-of-anemia-in-women-of-reprodutive-age.
2. Brimson S, Suwanwong Y, Brimson JM. Nutritional anemia predominant form of anemia in educated young Thai women. Ethn Health. 2019;24(4):405-14.
3. Khorwanichakij N, Kungwankiattichai S, Owattanapanich W. Validation of Several Formulas to Differentiate Thalassemia from Iron Deficiency Anemia and Proposal of a Thalassemia–Iron Deficiency Discrimination (TID) Predictive Score. Siriraj Medical Journal. 2022;74(4):256-65.
4. Akers AS, Howard D, Ford JC, editors. Distinguishing iron deficiency anaemia from thalassemia trait in clinical obstetric practice2018.
5. Jameel T, Baig M, Ahmed I, Hussain MB, Alkhamaly MBD. Differentiation of beta thalassemia trait from iron deficiency anemia by hematological indices. Pak J Med Sci. 2017;33(3):665-9.
6. Ferih K, Elsayed B, Elshoeibi AM, Elsabagh AA, Elhadary M, Soliman A, et al. Applications of Artificial Intelligence in Thalassemia: A Comprehensive Review. Diagnostics (Basel). 2023;13(9).

Early Version - Uncorrected

7. Anderson PG, Tarder-Stoll H, Alpaslan M, Keathley N, Levin DL, Venkatesh S, et al. Deep learning improves physician accuracy in the comprehensive detection of abnormalities on chest X-rays. Scientific Reports. 2024;14(1):25151.
8. Arora L, Singh SK, Kumar S, Gupta H, Alhalabi W, Arya V, et al. Ensemble deep learning and EfficientNet for accurate diagnosis of diabetic retinopathy. Scientific Reports. 2024;14(1):30554.
9. Brinker TJ, Hekler A, Enk AH, Berking C, Haferkamp S, Hauschild A, et al. Deep neural networks are superior to dermatologists in melanoma image classification. European Journal of Cancer. 2019;119:11-7.
10. Schömig-Markiefka B, Pryalukhin A, Hulla W, Bychkov A, Fukuoka J, Madabhushi A, et al. Quality control stress test for deep learning-based diagnostic model in digital pathology. Mod Pathol. 2021;34(12):2098-108.
11. Babker AM, Suliman RS, Ghazwani AAM, Alharbi W, Lyashenko VV. Combined Use of Edge Selection Operators in Individual Color Channels for the Analysis of Cytological Images Presented in RGB Format. Biomedical and Pharmacology Journal. 2025.
12. Babker AM, Suliman RS, Elshaikh RH, Boboyorov S, Lyashenko VV. Sequence of Simple Digital Technologies for Detection of Platelets in Medical Images. Biomedical and Pharmacology Journal. 2024.
13. Kansal K, Chandra TB, Singh A. ResNet-50 vs. EfficientNet-B0: Multi-Centric Classification of Various Lung Abnormalities Using Deep Learning. Procedia Computer Science. 2024;235:70-80.
14. Integration of EfficientNetB0 and Machine Learning for Fingerprint Classification. Informatica. 2023;47(5).
15. Chen X, Pu X, Chen Z, Li L, Zhao KN, Liu H, et al. Application of EfficientNet-B0 and GRU-based deep learning on classifying the colposcopy diagnosis of precancerous cervical lesions. Cancer Med. 2023;12(7):8690-9.
16. Ashames MMA, Demir A, Gerek ON, Fidan M, Gulmezoglu MB, Ergin S, et al. Are deep learning classification results obtained on CT scans fair and interpretable? Physical and Engineering Sciences in Medicine. 2024;47(3):967-79.

Early Version - Uncorrected

17. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making. 2019;19(1):281.
18. Shehab M, Abualigah L, Shambour Q, Abu-Hashem MA, Shambour MKY, Alsalibi AI, et al. Machine learning in medical applications: A review of state-of-the-art methods. Computers in Biology and Medicine. 2022;145:105458.
19. Uddin S, Lu H. Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data. PLOS ONE. 2024;19(4):e0301541.
20. Zhang R, Xu P, Guo L, Zhang Y, Li P, Yao D. Z-Score Linear Discriminant Analysis for EEG Based Brain-Computer Interfaces. PLOS ONE. 2013;8(9):e74433.
21. Stephens CR, Huerta HF, Linares AR. When is the Naive Bayes approximation not so naive? Machine Learning. 2018;107(2):397-441.
22. Wang N, Wang M, Zhou Y, Liu H, Wei L, Fei X, et al. Sequential Data-Based Patient Similarity Framework for Patient Outcome Prediction: Algorithm Development. J Med Internet Res. 2022;24(1):e30720.
23. Mavaie P, Holder L, Skinner MK. Hybrid deep learning approach to improve classification of low-volume high-dimensional data. BMC Bioinformatics. 2023;24(1):419.
24. Qaid TS, Mazaar H, Al-Shamri MYH, Alqahtani MS, Raweh AA, Alakwaa W. Hybrid Deep-Learning and Machine-Learning Models for Predicting COVID-19. Comput Intell Neurosci. 2021;2021:9996737.
25. Elshaikh RH, Amir R, Ahmeide A, Mohamedahmed KA, Alfeel AH, Higazi H, et al. Evaluation of the Discrimination between Beta-Thalassemia Trait and Iron Deficiency Anemia Using Different Indexes. International Journal of Biomedicine. 2022;12(3):375-9.

Early Version - Uncorrected

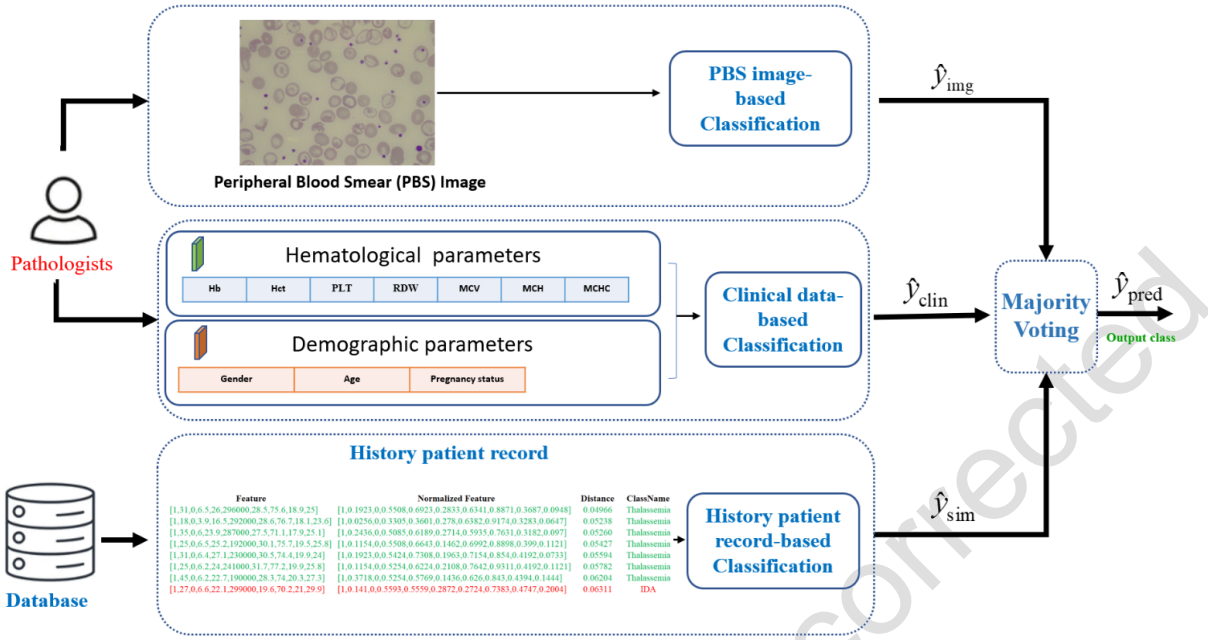


Figure 1. The overall proposed multimodal deep learning framework.

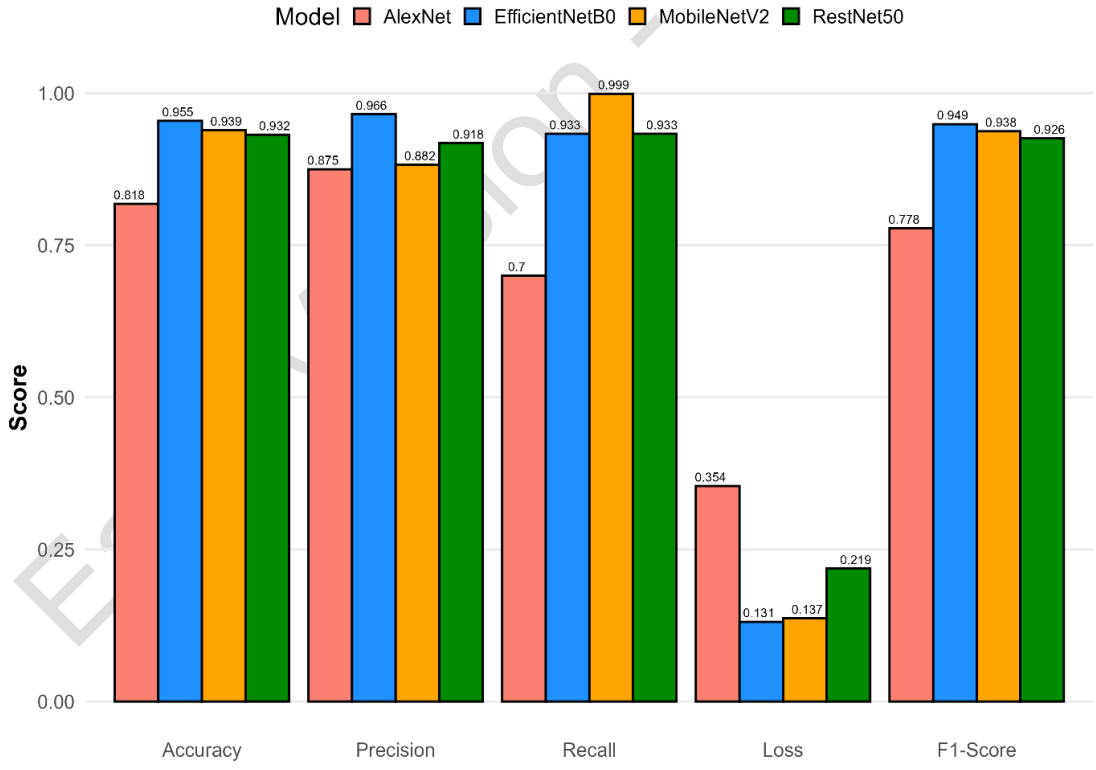


Figure 2. Performance of the four deep learning models.

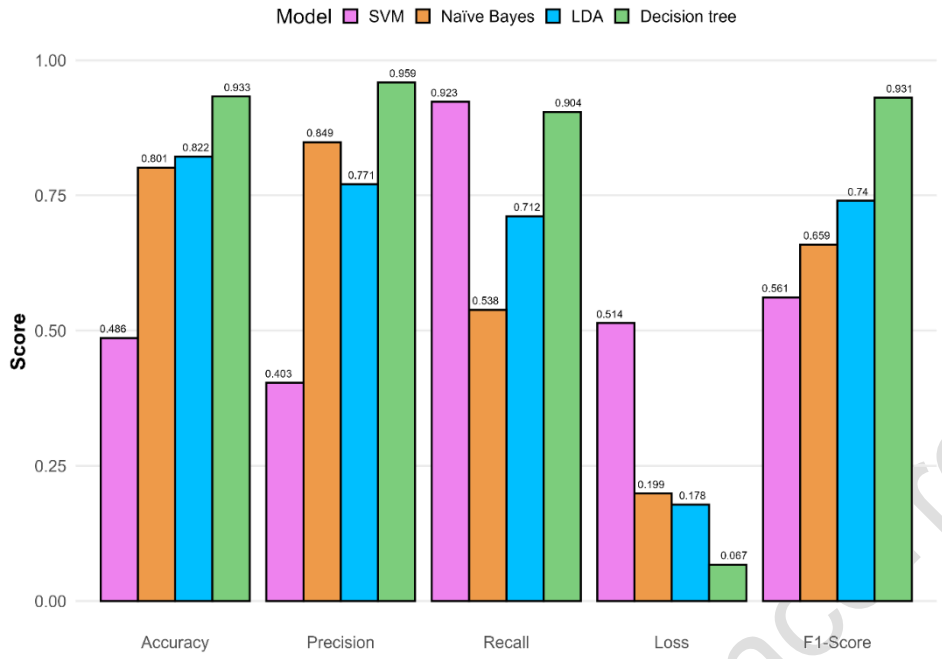


Figure 3. Performance of classification models.



Figure 4. Performance of KNN across different *k* values. X-axis is the *k* values, whereas Y-axis is the percentage.

Early Version - Uncorrected

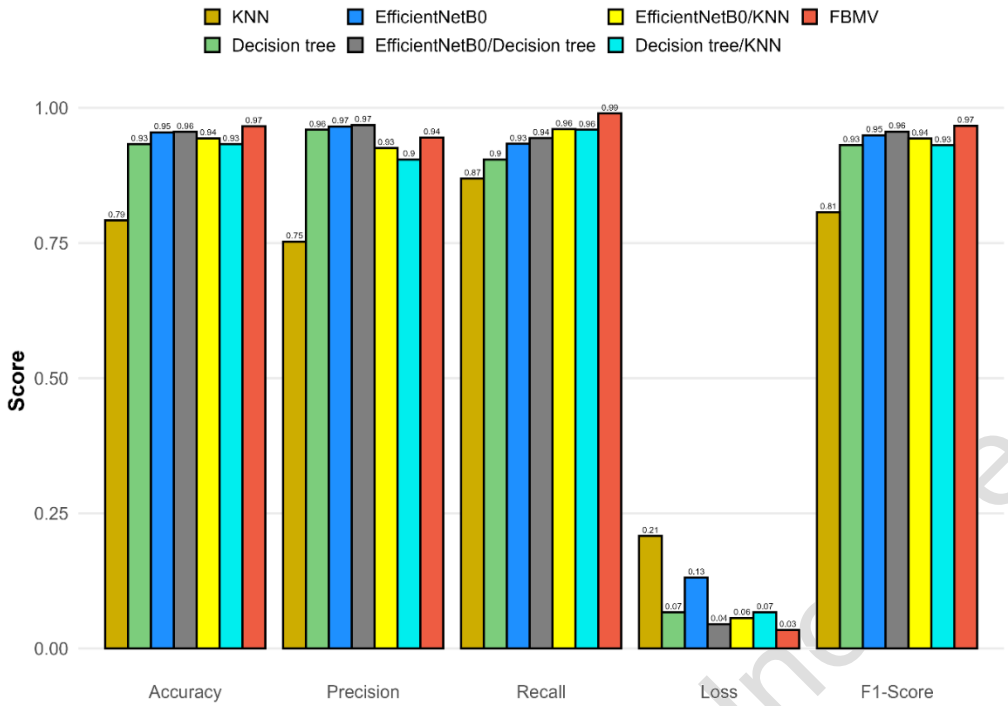


Figure 5. Performance of the fusion model following voting mechanisms.

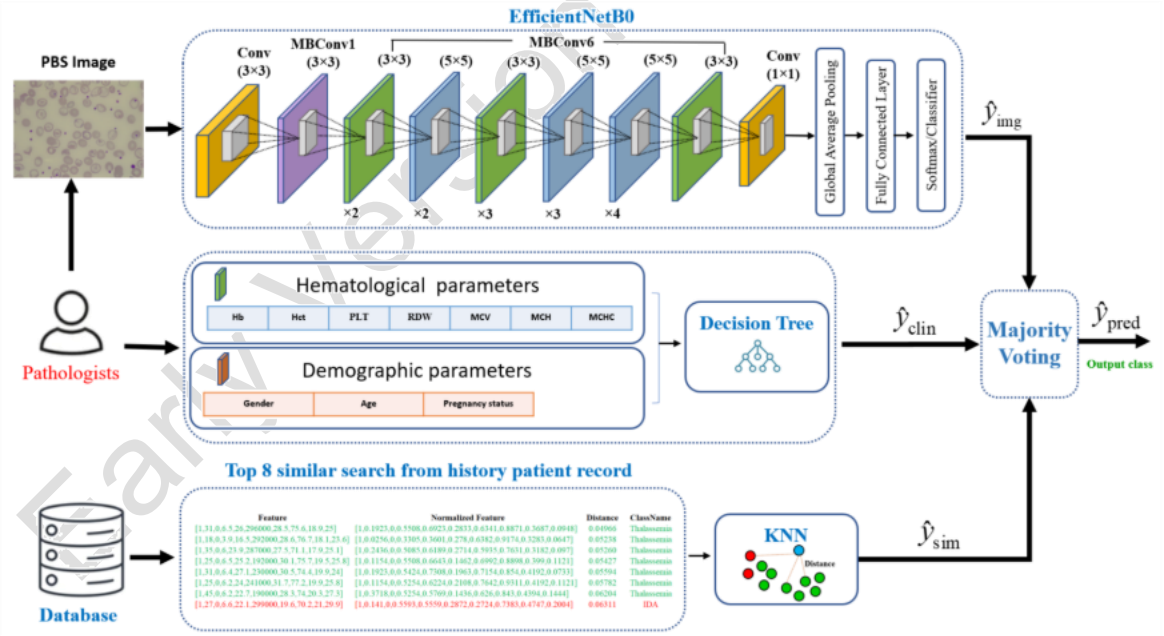


Figure 6. An illustration of the multimodal deep learning framework.

Table 1. Patient characteristics

	Female (N = 113)	**Male (N = 53)**	**Overall (N = 166)**
Age group			
Baby and adolescence	0 (0%)	0 (0%)	0 (0%)
Young Adults	46 (40.7%)	19 (35.8%)	65 (39.2%)
Middle-aged Adults	43 (38.1%)	21 (39.6%)	64 (38.6%)
Old Adults	24 (21.2%)	13 (24.5%)	37 (22.3%)
Pregnancy			
No	102 (90.3%)	53 (100%)	155 (93.4%)
Yes	11 (9.7%)	0 (0%)	11 (6.6%)
Hb typing			
A2F	1 (0.9%)	0 (0%)	1 (0.6%)
B trait	3 (2.7%)	1 (1.9%)	4 (2.4%)
E trait	8 (7.1%)	7 (13.2%)	15 (9.0%)
EA	1 (0.9%)	0 (0%)	1 (0.6%)
EA BART	2 (1.8%)	2 (3.8%)	4 (2.4%)
EE	10 (8.8%)	6 (11.3%)	16 (9.6%)
EF	27 (23.9%)	20 (37.7%)	47 (28.3%)
H	16 (14.2%)	4 (7.5%)	20 (12.0%)
H Cs	4 (3.5%)	2 (3.8%)	6 (3.6%)
IDA	41 (36.3%)	11 (20.8%)	52 (31.3%)
Hb			
Median [IQR]	7.60 [6.50,8.50]	7.90 [6.00,9.30]	7.70 [6.40,8.88]
Mean (SD)	7.52 (1.92)	7.75 (2.24)	7.59 (2.03)
Hct			
Median [IQR]	25.3 [22.0,27.7]	25.1 [20.0,28.9]	25.3 [21.4,28.0]
Mean (SD)	24.6 (5.22)	24.8 (6.72)	24.7 (5.72)

Early Version - Uncorrected

	Female (N = 113)	Male (N = 53)	Overall (N = 166)
WBC			
Median [IQR]	7020 [5690,9480]	9330 [7080,12800]	8020 [5890,10400]
Mean (SD)	8480 (4330)	10600 (4850)	9150 (4590)
Platelet			
Median [IQR]	276000 [207000,378000]	300000 [218000,376000]	282000 [211000,378000]
Mean (SD)	307000 (144000)	333000 (177000)	316000 (155000)
MCV			
Median [IQR]	64.0 [58.7,70.0]	64.0 [58.2,69.5]	64.0 [58.5,69.9]
Mean (SD)	64.5 (7.45)	63.3 (10.3)	64.1 (8.44)
MCH			
Median [IQR]	19.7 [17.6,21.5]	20.3 [17.4,23.5]	19.9 [17.5,21.9]
Mean (SD)	19.7 (3.14)	20.3 (4.06)	19.9 (3.46)
MCHC			
Median [IQR]	31.2 [28.0,32.7]	31.6 [29.3,33.0]	31.3 [28.4,32.8]
Mean (SD)	30.6 (4.57)	31.0 (2.74)	30.7 (4.07)
RDW CV			
Median [IQR]	21.9 [18.6,26.5]	23.9 [21.3,27.0]	22.4 [18.7,26.8]
Mean (SD)	22.6 (5.54)	23.9 (5.71)	23.0 (5.61)
Ferritin			
Median [IQR]	258 [14.0,737]	679 [137,1540]	381 [22.3,958]
Mean (SD)	890 (1870)	1080 (1320)	950 (1710)
Blood transfusion			
No	70 (61.9%)	33 (62.3%)	103 (62.0%)
Yes	43 (38.1%)	20 (37.7%)	63 (38.0%)
Folic supplement			
No	45 (39.8%)	24 (45.3%)	69 (41.6%)
Yes	68 (60.2%)	29 (54.7%)	97 (58.4%)

Early Version - Uncorrected

	Female (N = 113)	Male (N = 53)	Overall (N = 166)
Iron supplement			
No	91 (80.5%)	46 (86.8%)	137 (82.5%)
Yes	22 (19.5%)	7 (13.2%)	29 (17.5%)

Abbreviations: Hb = hemoglobin; Hct = hematocrit; WBC = white blood cell count; MCV = mean corpuscular volume; MCH = mean corpuscular haemoglobin; MCHC = mean corpuscular haemoglobin concentration; RDW = red cell distribution width; IDA = iron deficiency anemia; HbA2F = normal hemoglobin pattern; B trait = β-thalassemia trait; E trait = hemoglobin E trait; EA = hemoglobin E with hemoglobin A; EA BART = hemoglobin E with Hb Bart's; EE = homozygous hemoglobin E; EF = hemoglobin E with hemoglobin F; H = hemoglobin H disease; H Cs = hemoglobin H with Constant Spring.

Supplementary Data

Supplementary Table S1. Subgroup analysis evaluating model performance across demographic strata (sex and age groups).

Group	Accuracy	Precision	Recall	F1 score
Male	0.96	0.94	0.99	0.97
Female	0.90	0.91	0.91	0.91
Young adults	0.95	0.99	0.90	0.95
Middle-aged adults	0.87	0.82	0.95	0.88
Older adults	0.96	0.99	0.93	0.96

Early Version - Uncorrected

Supplementary Table S2. Performance metrics of the evaluated deep learning models obtained using patient-wise data separation.

Model	Accuracy	Precision	Recall	F1-score	Loss	ROC AUC	PR AUC
AlexNet	0.615	0.600	0.500	0.545	0.578	0.762	0.571
EfficientNetB0	0.846	0.833	0.833	0.833	0.490	0.833	0.686
MobileNetV2	0.769	0.667	0.990	0.800	0.347	0.929	0.758
ResNet50	0.538	0.500	0.500	0.500	0.823	0.619	0.560

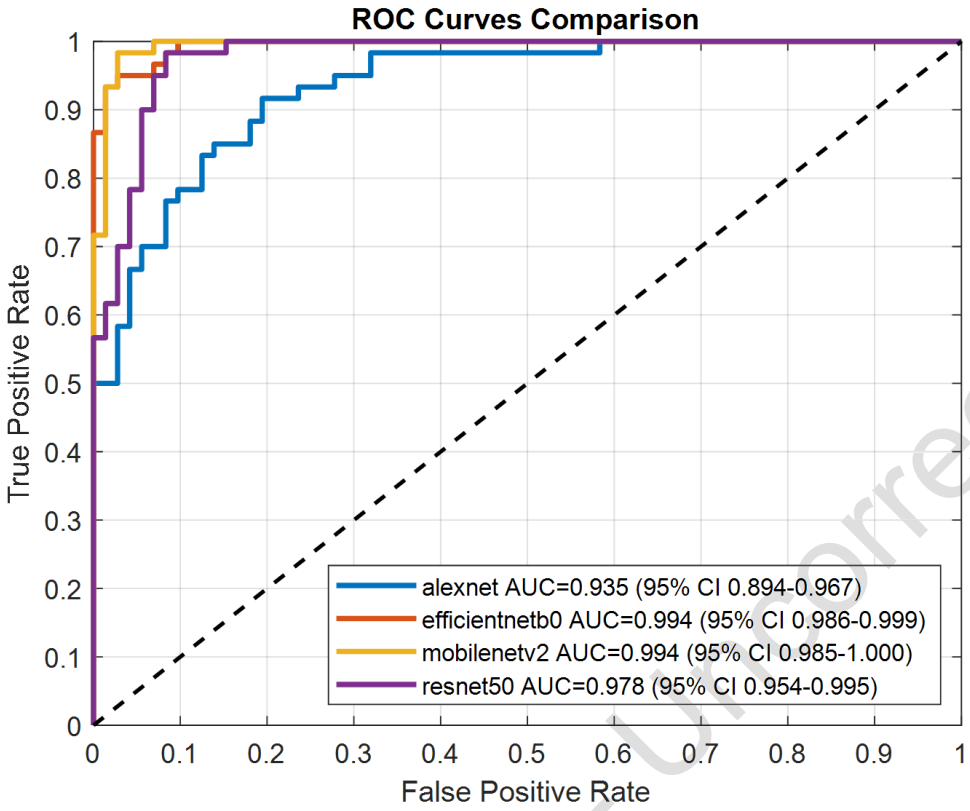
Supplementary Table S3. Pruning level and their decision tree performance.

Level	Resubstitution error at each pruning level	Accuracy	Recall	Precision	F1-score	Loss
0	0.0411	0.95890	0.92308	0.96000	0.94118	0.041096
1	0.0616	0.93836	0.90385	0.92157	0.91262	0.061644
2	0.0890	0.91096	0.86538	0.88235	0.87379	0.089041
3	0.1096	0.89041	0.78846	0.89130	0.83673	0.109590
4	0.1712	0.82877	0.88462	0.70769	0.78632	0.141230

Supplementary Table S4. Sensitivity analysis of distance metrics used in the KNN similarity layer.

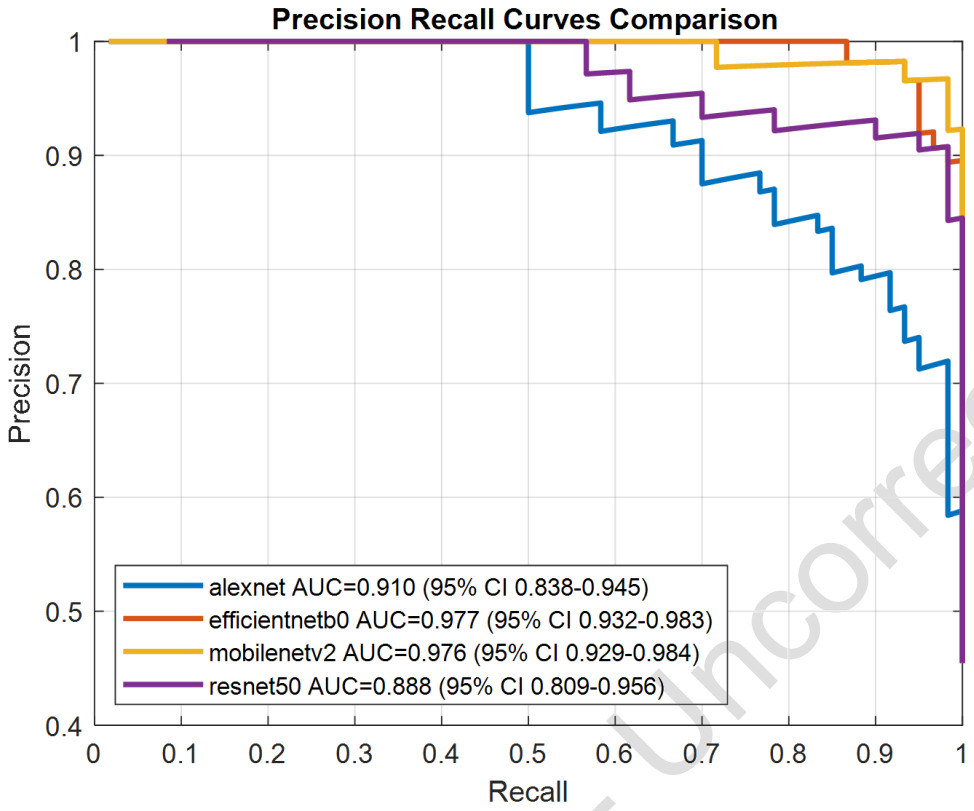
Distance metric	Accuracy
Euclidean	0.83562
Cosine	0.82877
Manhattan	0.82192
Mahalanobis	0.82192

Early Version - Uncorrected



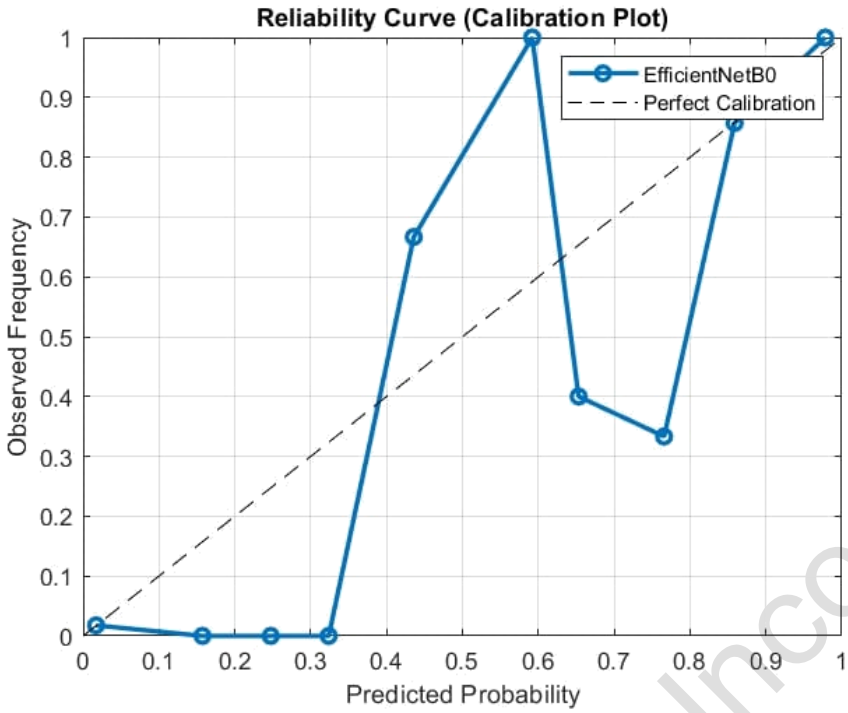
Supplementary Figure S1. Receiver operating characteristic (ROC) curves of the evaluated deep learning models based on image-level predictions.

Early Version - Uncorrected



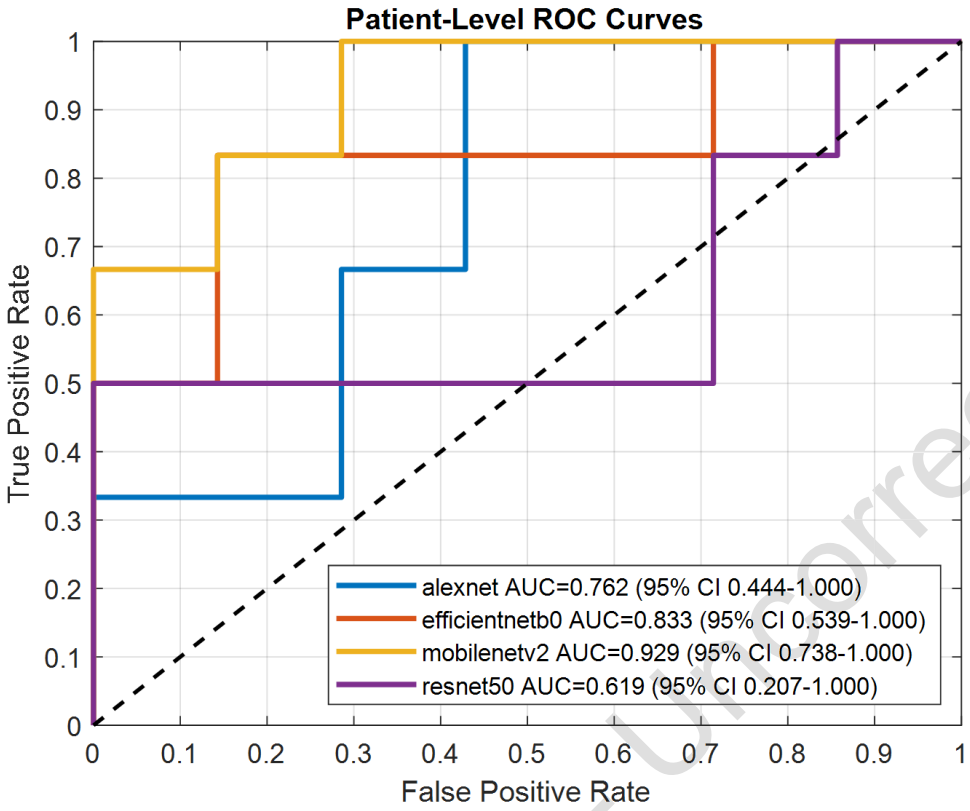
Supplementary Figure S2. Precision–recall (PR) curves of the evaluated deep learning models based on image-level predictions.

Early Version - Uncorrected



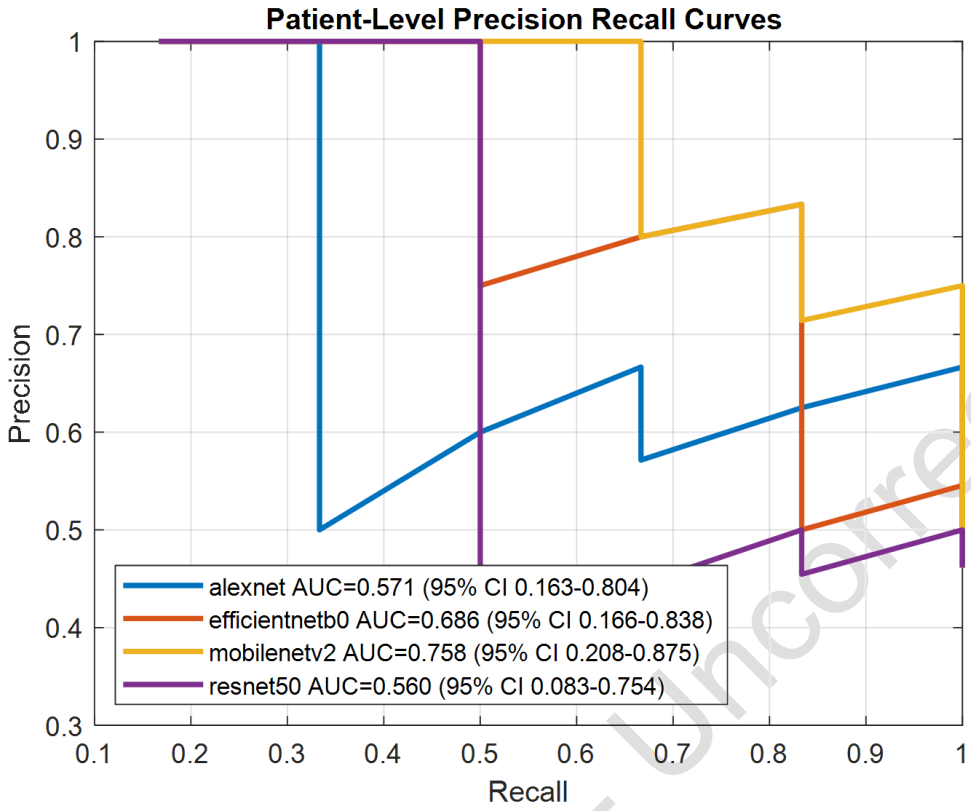
Supplementary Figure S3. Reliability curve of EfficientNetB0 showing the relationship between predicted probabilities and observed outcome frequencies. The dashed line represents perfect calibration.

Early Version - Uncorrected



Supplementary Figure S4. Receiver operating characteristic (ROC) curves obtained using patient-wise data separation for the evaluated deep learning models, confirming consistent model performance without potential image-level overlap.

Early Version - Uncorrected



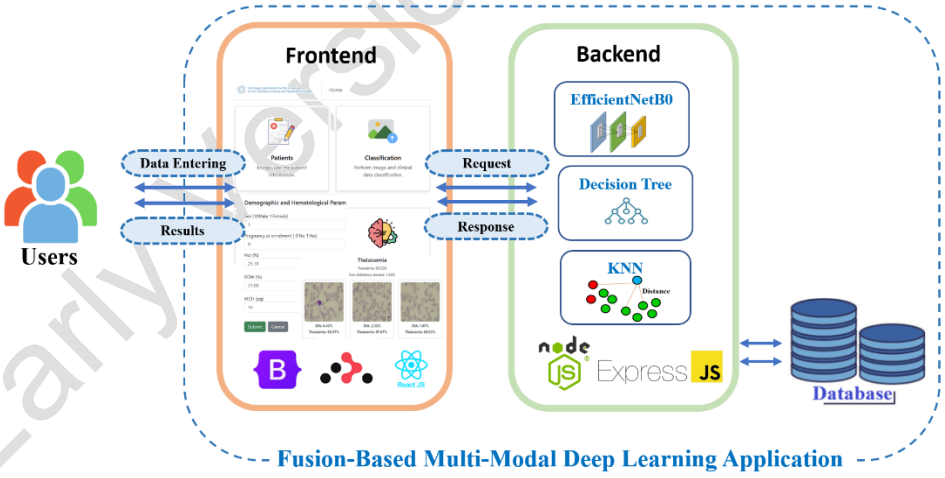
Supplementary Figure S5. Precision–recall (PR) curves of the evaluated deep learning models using patient-wise data separation, confirming consistent model performance without potential image-level overlap.

Early Version - Uncorrected

Supplementary Technical Information

Web Application Architecture

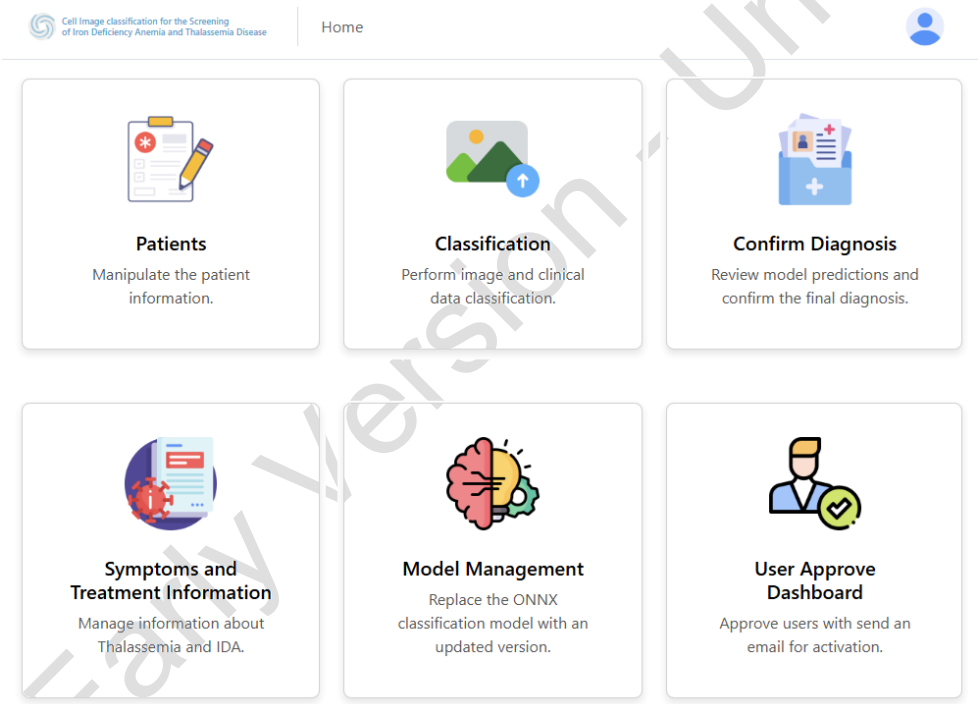
A web-based application was developed to facilitate interaction between users and the proposed framework. The system architecture comprises two main components: the frontend (client side) and the backend (server side). The frontend is implemented using React, with React Router and Bootstrap integrated to ensure a responsive interface and efficient navigation. The backend is developed using Node.js and Express, connected to a MySQL database for secure data management. The EfficientNetB0 model was trained using the Adam optimizer with a mini-batch size of 10 and a maximum of 6 epochs. The initial learning rate was set to 1×10^{-4}, and the data were shuffled at every epoch. Prior to training, data augmentation was applied with a random scaling factor in the range of 0.9–1.1. Model performance was evaluated using accuracy as the metric, with validation performed every three iterations on the augmented validation dataset. To enable deployment in resource-constrained environments, the trained model was exported to the Open Neural Network Exchange (ONNX) format using the MATLAB command. The decision tree model was exported in JavaScript Object Notation (JSON) format. The constructed tree was fully grown without pruning and reached a maximum depth of six. On the server side, the EfficientNetB0, decision tree, and KNN inference modules are integrated to provide decisions, as shown in **Supplementary Figure S6,** which depicts the architecture of the proposed fusion-based multi-modal deep learning web application.



Supplementary Figure S6. Architecture of the proposed fusion-based multi-modal deep learning application.

Early Version - Uncorrected

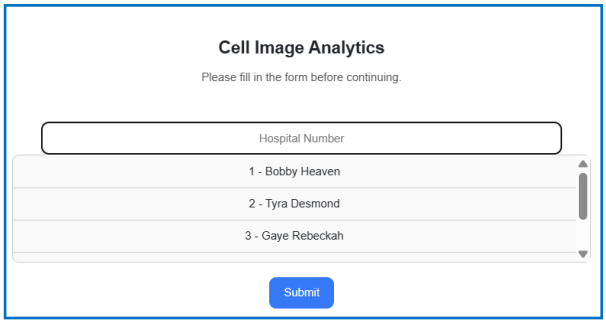
The graphical user interface (GUI) adopts a clean, card-based layout to facilitate intuitive use. All users must authenticate via a secure login page before accessing the system. Once the pathologist logs into the system, six interactive cards are presented, each corresponding to a core system function. The *Patients* card provides structured forms for adding, viewing, and updating patient information. The *Classification* interface enables the upload of PBS images and entry of demographic and hematological data. The *Confirm Diagnosis* panel allows physicians to review the model predictions and confirm the diagnosis. The *Symptoms and Treatment Information* feature presents user-friendly educational content on thalassemia and IDA. The *Model Management* utility facilitates the replacement of the model with an updated version. Lastly, the *User Approval* dashboard offers administrative tools for reviewing and activating new user accounts via email. **Supplementary Figure S7** shows the application interface for the pathologist.



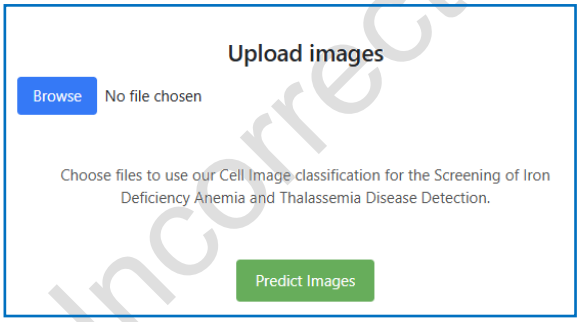
Supplementary Figure S7. Graphical user interface showing the functional components accessible to the pathologist.

Early Version - Uncorrected

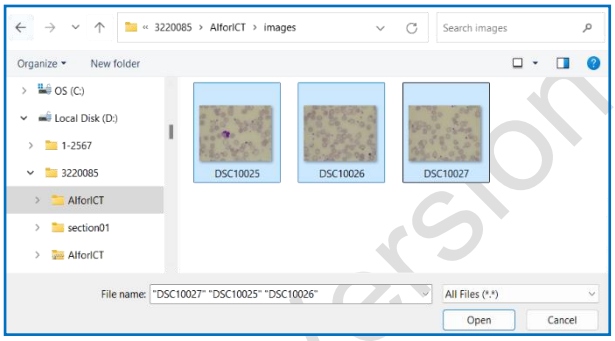
Supplementary Figure S8 illustrates the workflow of the image-based classification module. The pathologist first inputs the patient's Hospital Number (HN) (a). Next, the upload interface allows image selection (b), followed by the file selection dialog supporting multiple PBS images (c). Upon clicking the *Predict Images* button, the images are resized and processed by the ONNX model, which estimates the probability of thalassemia or IDA. The results are then aggregated to determine the classification outcome (d).



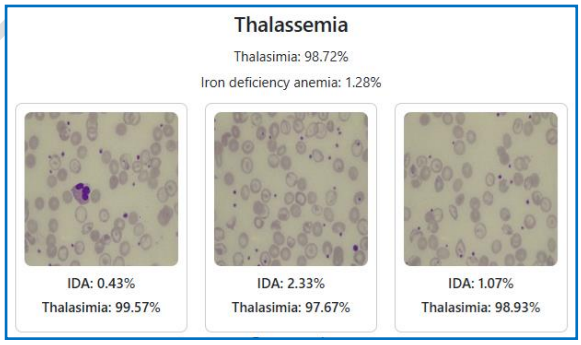
(a) HN selection screen



(b) Image upload interface



(c) File selection dialog

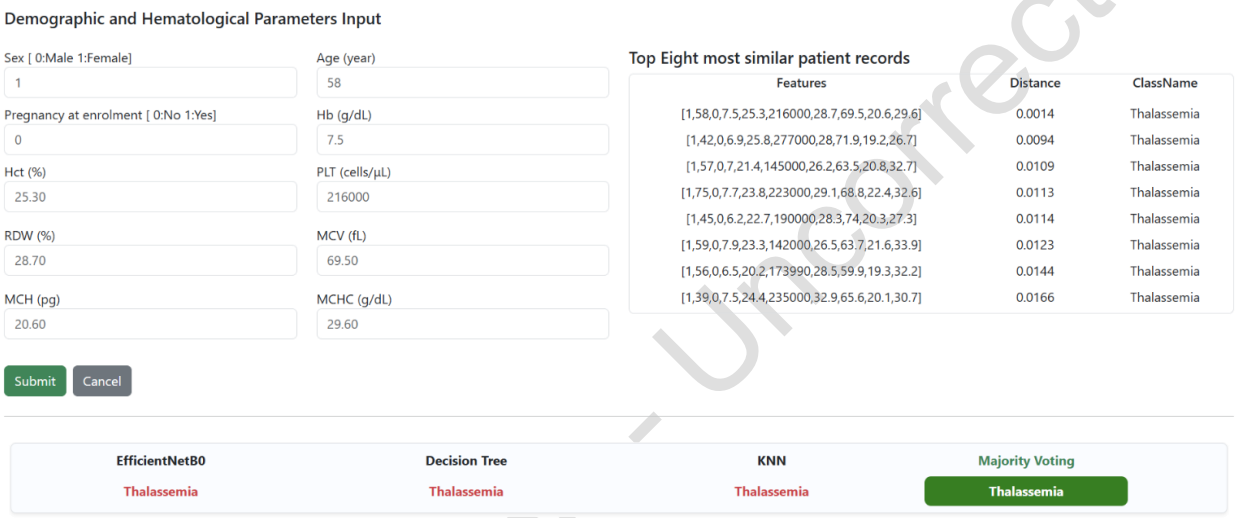


(d) Classification results display

Supplementary Figure S8. PBS image-based classification workflow. (a) HN selection, (b) image upload interface, (c) file selection dialog, (d) classification results display.

Early Version - Uncorrected

Supplementary Figure S9 shows the interface for entering demographic and hematological data. For the given example (Sex = 1, Age = 58, Pregnancy status = 0, Hb = 7.50 g/dL, Hct = 25.30%, PLT = 216,000 cells/µL, RDW = 28.70%, MCV = 69.50 fL, MCH = 20.60 pg, MCHC = 29.60 g/dL), both the decision tree and KNN predict thalassemia. The application also retrieves and displays the eight most similar patient records. Finally, predictions from the EfficientNetB0, decision tree, and KNN are combined via the majority voting process, producing a final FBMV classification of thalassemia.



Supplementary Figure S9. Example input of ten demographic and hematological parameters (Sex = 1, Age = 58, Pregnancy status = 0, Hb = 7.50, Hct = 25.30, PLT = 216,000, RDW = 28.70, MCV = 69.50, MCH = 20.60, MCHC = 29.60). All three models (EfficientNetB0, decision tree, and KNN) predict thalassemia, and the majority voting scheme yields a final classification of thalassemia.

Early Version - Uncorrected

Supplementary Figure S10 shows another example of demographic and hematological data. In this case (Sex = 1, Age = 70, Pregnancy status = 0, Hb = 7.10 g/dL, Hct = 22.90%, PLT = 376,000 cells/µL, RDW = 17.00%, MCV = 72.90 fL, MCH = 22.60 pg, MCHC = 31.00 g/dL), both the decision tree and KNN predict IDA. Although EfficientNetB0 predicts thalassemia, the majority voting scheme yields a final classification of IDA, since two of the three models predict this class.



Supplementary Figure S10. Example input of ten demographic and hematological parameters (Sex = 1, Age = 70, Pregnancy status = 0, Hb = 7.10, Hct = 22.90, PLT = 376,000, RDW = 17.00, MCV = 72.90, MCH = 22.60, MCHC = 31.00). The decision tree and KNN models predict IDA, while EfficientNetB0 predicts thalassemia. Majority voting yields a final classification of IDA.

Early Version Uncorrected