Taylor and Francis Book Chapter

Contents

CHAPTER 1 ▪ A Semi-Supervised and Explainable Stacked Ensemble Model for Multi-Class Haematological Disorder Classification Using Complete Blood Count Data 1

KOPIL DAS, NOSHIN AMINA, MD ABDUL MUTALIB, NASRIN CHOWDHURY , and MOHAIMINUL ISLAM

1.1 INTRODUCTION 2

1.2 RELATED WORK 4

1.3 METHODOLOGY 5

1.3.1 Datasets 5

1.3.2 Feature Engineering 6

1.3.3 Preprocessing: Imputation and Scaling 7

1.3.4 Semi-Supervised Learning with Label Spreading 7

1.3.5 Classification Model: Stacked Ensemble with Seed Averaging 7

1.3.6 Model Interpretation and XAI Evaluation 7

1.4 RESULTS AND EVALUATION 8

1.4.1 Data Sets, Preprocessing, and Semi-Supervised Label Spreading 8

1.4.2 Ensemble Classifier and Hyperparameter Tuning 8

1.4.3 Evaluation Metrics 8

1.4.4 Quantitative Results 9

1.4.5 Comparative Analysis 12

1.4.6 Ablation Study 13

1.5 CONCLUSION AND FUTURE WORK 13

CHAPTER 1

A Semi-Supervised and Explainable Stacked Ensemble Model for Multi-Class Haematological Disorder Classification Using Complete Blood Count Data

Kopil Das

North East University Bangladesh, Dept. of CSE, Sylhet 3100, Bangladesh.
Email: kopildas.cs@gmail.com

Noshin Amina

Comilla University, Dept. of CSE, Comilla 3506, Bangladesh.
Email: noshinamina24@gmail.com

Md Abdul Mutalib

Harbin Institute of Technology, Dept. of Computer Science and Technology, Shenzhen 518055, China.
Email: 25sf51110@stu.hit.edu.cn; mutalib.cs@gmail.com

Nasrin Chowdhury

Norwegian University of Science and Technology, Dept. of Clinical and Molecular Medicine, Trondheim 7034, Norway.
Email: nasrin.ntnu@gmail.com

Mohaiminul Islam

Simula Research Laboratory, Oslo 0164, Norway.
Email: mohaiminul100@hotmail.com

CONTENTS

1.1 Introduction 2
1.2 Related Work 4
1.3 Methodology 5
1.3.1 Datasets 5
1.3.2 Feature Engineering 6
1.3.3 Preprocessing: Imputation and Scaling 6
1.3.4 Semi-Supervised Learning with Label Spreading 7
1.3.5 Classification Model: Stacked Ensemble with Seed Averaging ... 7
1.3.6 Model Interpretation and XAI Evaluation 7
1.4 Results and Evaluation 8
1.4.1 Data Sets, Preprocessing, and Semi-Supervised Label Spreading 8
1.4.2 Ensemble Classifier and Hyperparameter Tuning 8
1.4.3 Evaluation Metrics 8
1.4.4 Quantitative Results 9
1.4.5 Comparative Analysis 12
1.4.6 Ablation Study 13
1.5 Conclusion and Future Work 13

Accurate interpretation of Complete Blood Count (CBC) data is essential for diagnosing haematological disorders, but manual analysis is time-consuming, subjective, and challenging in resource-constrained clinical settings. This research proposes a semi-supervised multi-class classification method for haematological diseases based on regular CBC values, clinically relevant feature engineering, a seed-averaged stacked ensemble classifier, and explainable artificial intelligence approaches.

The proposed architecture is built on 1,281 labeled CBC samples at the patient level and 643 unlabeled routine CBC reports. We trained a stacked ensemble of Random Forest, Extra Trees, Support Vector Classifier, Logistic Regression, Gradient Boosting and XGBoost with logistic regression as a meta-learner. Training data was resampled to handle class imbalance using SMOTE Tomek resampling. The final seed-averaged model attained the accuracy of 94.94% on the held-out test set.

Permutation importance and SHAP explainability analysis showed clinically relevant predictors such as haemoglobin level, platelet count, white blood cell count, MCV-related characteristics and red-cell indices. The results suggest that semi-supervised learning, engineered CBC features and stacked ensemble modeling can give an effective and interpretable solution for deployment in low-resource clinical environments to multi-class haematological disorder classification utilizing routine CBC data.

1.1 INTRODUCTION

Anemia, leukemia, and thrombocytopenia are common hematological disorders that create a major global health burden. These conditions may lead to serious complications when diagnosis is delayed or inaccurate. Routine Complete Blood Count (CBC) testing is one of the most widely used and cost-effective laboratory investigations for

early screening and clinical assessment. CBC measures provide quantitative information on circulating blood cells, including red blood cell count (RBC), hemoglobin (HGB), hematocrit (HCT), platelet count (PLT), and white blood cell (WBC) differentials [1]. These parameters are generally an indication of serious pathological changes and are frequently used for the diagnosis of anemia, infections, leukemia and abnormalities of platelets. However, manual interpretation of CBC findings can be time consuming, subjective, and prone to errors, particularly in busy hospitals and resource-limited settings where the availability of specialist expertise may not always be guaranteed [2].

Recent progress in artificial intelligence (AI) and machine learning (ML) provides novel possibilities for the automation of hematological diagnostics and the support of clinical decision-making. Several studies have shown that ML models can effectively identify anemia, screen for leukemia, and classify blood disorders based on CBC data [3, 4, 9, 12, 13]. Population scale studies have also demonstrated that CBC based models can generalize across diverse cohorts. For example, Pullakhandam and McCoy constructed an explainable model for iron deficiency anemia with more than 19,000 samples and external validation data [5]. Schipper et al. also demonstrate strong predicting performance for several hemoglobinopathies with conventional CBC parameters on external cohorts [8]. These results demonstrate the practical usefulness of everyday CBC data for AI-assisted diagnosis.

Despite these positive outcomes, a number of challenges remain. Most existing research focus primarily on binary classification tasks such as anemia vs non-anemia or iron deficiency anemia vs thalassemia [7, 16]. Some techniques require extra laboratory parameters beyond regular CBC testing, and others show lower robustness to class imbalance or heterogeneous clinical data [11, 10]. In addition, many high-performing black-box models provide limited interpretability, which may reduce clinician trust and slow real-world adoption. Reliable probability calibration and multi-class diagnosis are also less frequently explored in prior work [2].

To address these challenges, this study proposes a semi-supervised multi-class hematological diagnosis framework using only routine CBC data. The proposed pipeline utilizes both labeled (1,281 samples) and unlabeled (643 samples) datasets to improve learning efficiency and data utilization. Clinically meaningful ratios, indicator flags, and thalassemia-related indices were engineered to strengthen diagnostic representation. A stacked ensemble classifier was then developed to combine the strengths of multiple machine learning models for more robust prediction. Experimental results show that the proposed model achieved 94.94% accuracy with a macro ROC–AUC of 0.98 on the held-out test set. Furthermore, SHAP-based explanations were used to identify influential CBC features and improve model transparency, supporting clinically interpretable decision-making.

Contributions include:

- A multi-class semi-supervised framework combining confidence-based label spreading, and ensemble learning.
- CBC indices, anemia indicators, inflammatory ratios, and thalassemia screening indices are utilize for clinically interpretable feature engineering.

- An extensive XAI assessment employing permutation importance, SHAP analysis, seed-stability testing, and reliability/deletion evaluations.

1.2 RELATED WORK

Machine learning approaches using CBC data have shown strong performance for anemia and hematological disorder classification [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. Ensemble methods, SHAP interpretability, and gradient-boosted models have been employed for high accuracy. However, most studies focus on binary or limited multi-class problems, neglect unlabeled real-world CBC data, and often ignore probability calibration.

Airlangga [3] analyzed various algorithms such as decision trees, random forests, XGBoost, support vector machines, logistic regression and k-nearest neighbors and found that a simple decision tree produced the highest balanced accuracy of 94.17% in a five-fold cross-validation. This shows that CBC features themselves provide a lot of information for diagnosis even without using very complicated models. Similar conclusions were reached by Kumar et al. [4], who showed that decision-tree-based models may stay competitive for anemia classification when combined with effective feature selection [16].

Large population-based studies further confirmed the effectiveness of CBC-driven machine learning. Pullakhandam and McRoy [5] developed an explainable gradient-boosted model for iron-deficiency anemia (IDA) using more than 19,000 samples from the NHANES database and validated it on an external Kenyan cohort. The model produces high recall values (0.98 and 0.89 for the US and Kenyan cohorts, respectively), suggesting robust generalization across populations. Lv et al. [6] explored a smaller cohort study of 193 patients with microcytic hypochromic anemia and reported AUC values around 0.98 for random forest and logistic regression models, however, the small size of the data limited the robustness of the findings [17].

Differentiation of IDA from thalassemia has been one of the key scientific interests on CBC. This is a clinical challenge because of the similar hematological findings. Tepakhan et al. [7] studied 1,143 cases of microcytic anemia with random forest and gradient boosting classifiers and achieved 90.7% accuracy for binary IDA vs thalassemia classification. However, when expanded to the three-class case, the performance drops dramatically, showing the inherent difficulties in multi-class diagnosis. Similar conclusions were observed elsewhere in other anemic investigations, where high accuracy in binary situations did not transfer well to multi-class tasks [18]. In another large multicenter study, Schipper et al. [8] showed that XGBoost and logistic regression models trained with standard CBC parameters achieved AUROC values of 0.95 to 0.98 for multiple hemoglobinopathies in external cohorts, demonstrating the scalability of CBC-based models when validated across diverse populations [18].

Beyond anemia, CBC parameters have also been applied to leukemia detection and classification. Cheng et al. [9] analyzed pediatric CBC data and reported AUC values of 0.950 for leukemia subtype prediction and 0.909 for distinguishing acute myeloid leukemia from other conditions using a random forest model [4]. Other studies using routine CBC data for leukemia screening reported similarly encouraging results

but noted reduced performance under class imbalance and dataset heterogeneity [10, 11]. These findings suggest that while CBC-based models are effective, robustness across clinical settings remains a challenge.

Several works have proposed ensemble learning approaches to improve diagnostic accuracy and stability. Pacal [12] and Cakmak [13] reported accuracies exceeding 98% using Gradient Boosting, LightGBM, and CatBoost models for anemia and blood disorder classification. Hindi [14] demonstrated that stacking ensembles outperform individual classifiers, particularly for handling class imbalance and rare anemia subtypes [19]. However, most ensemble-based studies rely exclusively on supervised learning and require fully labeled datasets, which may not reflect real-world clinical conditions where unlabeled CBC reports are common.

Explainable AI methods have increasingly been incorporated into CBC-based diagnostic models to improve clinical trust and interpretability. Techniques such as SHAP have been used to relate model predictions to physiologically meaningful CBC parameters, supporting clinical validation of machine learning outputs [5, 15]. Nevertheless, many existing studies do not systematically evaluate probability calibration, despite its importance for risk-aware clinical decision support systems.

Existing literature shows that CBC-based machine learning models can achieve high accuracy for specific hematological diagnosis tasks, particularly in binary classification settings. However, we observe several important limitations across prior studies, including limited focus on multi-class hematological diagnosis, underutilization of unlabeled real-world CBC data, insufficient attention to probability calibration, and a lack of hybrid approaches that integrate clinical knowledge with data-driven learning [2].

In our work, we address these gaps by proposing a semi-supervised, ensemble-based framework that combines domain-informed feature engineering, interpretable modeling, calibrated prediction and post-hoc explainability. Our approach aims to enable accurate, scalable, and clinically reliable multi-class hematological diagnosis using unlabeled CBC data.

1.3 METHODOLOGY

The pipeline includes preprocessing and feature engineering, semi-supervised learning, stacked ensemble classification, and evaluation (Fig. 1.1).

1.3.1 Datasets

This study used two datasets: a labeled public CBC dataset and an unlabeled real-world CBC dataset. The labeled dataset contained 1,281 samples with 14 standard Complete Blood Count (CBC) parameters and nine diagnostic classes related to different haematological disorders. This dataset was used for model training and final evaluation.

The unlabeled dataset included 654 routine CBC reports gathered from Sylhet MAG Osmani Medical College, Bangladesh. In total, the unlabeled dataset has 10,994 items at the parameter level. For 643 reports, most of them of the usual CBC data were available including WBC, RBC, platelet count, PCV/HCT, MCV, MCH,

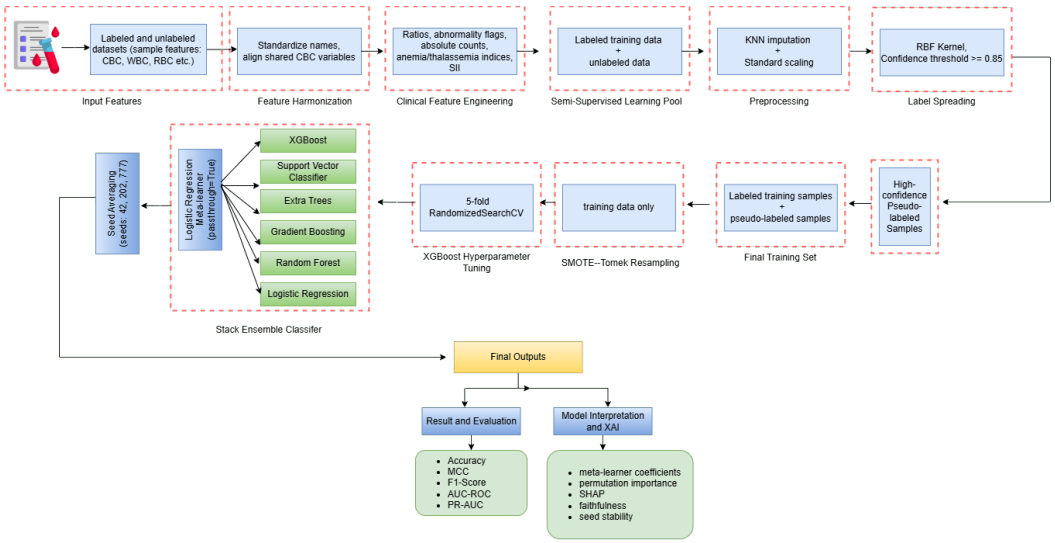


Figure 1.1 The proposed model architecture illustrates the complete workflow from preprocessing to final prediction.

MCHC, RDW-CV, RDW-SD, MPV and WBC differentials. These data were used in the semi-supervised learning stage to improve the learning process by utilizing additional unlabeled samples. Before combining the datasets, feature names were standardized to maintain consistency. Features with more than 50% missing values were removed, while the remaining missing values were handled during preprocessing.

We separated the labeled dataset into 80% training and 20% held-out testing for evaluation using stratified sampling. In semi-supervised learning the unlabeled data were used only with the training subset, while the test set was fully inaccessible.

1.3.2 Feature Engineering

To integrate clinically meaningful structure into both labeled and unlabeled CBC data, we applied the same feature-engineering procedure. We calculated inflammation-related ratios, including the neutrophil-to-lymphocyte ratio (NLR) and the platelet-to-lymphocyte ratio (PLR) [20, 21], and incorporated standard clinical indicators for anemia (reduced hemoglobin), microcytosis (MCV $<$ 80fL), macrocytosis (MCV $>$ 100fL), and unusual red blood cell count. We also calculated absolute neutrophil and lymphocyte counts from white blood cell (WBC) and differential percentages and incorporated commonly used anemia/thalassemia screening indices (Mentzer, Green & King, England & Fraser, Shine & Lal, and Srivastava) [22, 23, 24, 25], in addition to the systemic immune-inflammation index (SII) [26] and an elevated red cell distribution width (RDW) indicator. These derived features convert conventional hematological principles into numerical representations, which enhance both predicted accuracy and interpretability.

1.3.3 Preprocessing: Imputation and Scaling

Missing CBC values were imputed utilizing a distance-weighted `KNNImputer` ($k = 5$) trained on the SSL training pool (which included labeled and unlabeled data). Then, features were normalized using `StandardScaler`. Both transformers were trained on the training pool and subsequently applied unaltered to the SSL pool and the held-out test set.

1.3.4 Semi-Supervised Learning with Label Spreading

In order to use unlabeled CBC data, we integrated the labeled training set with the unlabeled data to create a semi-supervised learning pool, preprocessed it with the established imputer and scaler, and assigned an initial label -1 to the unlabeled samples. After that, we applied `LabelSpreading` (RBF kernel, $\alpha = 0.2$, max_iterations $= 1000$) to propagate labels throughout the system and derive class posteriors. Unlabeled samples were assigned pseudo-labels only when the maximum class probability surpassed 0.85, and the remaining samples were eliminated to reduce label noise. The final training set consequently consisted of the original labeled training instances in addition to these high-confidence pseudo-labeled samples. To reduce the class imbalance and minimize overlap, we employed SMOTE and Tomek resampling before training the final classifiers.

1.3.5 Classification Model: Stacked Ensemble with Seed Averaging

We utilized a stacked ensemble to integrate different classifiers. The foundational layer comprised Random Forest (800 trees), Extra Trees (400 trees), an RBF-kernel SVC with probability outputs, multinomial Logistic Regression (all utilizing `class_weight="balanced"` where relevant), Gradient Boosting, and an XGBoost model optimized via `RandomizedSearchCV`. A logistic regression meta-learner was trained with 5-fold out of fold predictions with `passthrough=True`, enabling the use of both base-model probabilities and novel features. Class imbalance was addressed using SMOTE-Tomek on the training dataset, then refitting the scaler on the resampled set. To reduce variance, the complete stacking technique was done using three random seeds (42, 202, 777), and the final predictions were derived by averaging class probabilities across the seeds.

1.3.6 Model Interpretation and XAI Evaluation

The interpretability of the models was examined using complementary global and class-specific studies. First, we considered the coefficients of the multinomial logistic-regression meta-learner in the stacked ensemble. The meta-learner was presented with the base-model probability outputs and engineered CBC features by because of `passthrough=True`. Mean absolute coefficients across classes were consequently used to identify influential meta-features.

Second, final-model permutation importance was computed for the test set to quantify the effect of each feature on predictive performance. Seed-specific stacked

models were used to produce importance scores which were pooled to derive stable global feature rankings.

Third, SHAP analysis was applied to the fitted XGBoost component of the stack. The mean absolute SHAP values were summarized worldwide and per class to explore whether the model depended on clinically important CBC markers such as HGB, PLT, WBC, MCV, MCH, MCHC, and derived hematological indices.

Finally, we examined the explanation reliability by the fidelity test and stability tests. For the fidelity study, the top-k essential features were perturbed and the increase in log loss was compared to a random-$k$perturbation baseline. Stability was evaluated by comparing feature importance rankings across random seeds using Spearman correlation and top-k overlap. These studies were utilized to confirm that the explanations were both clinically meaningful and consistent across repeated training runs.

1.4 RESULTS AND EVALUATION

"nobreak

1.4.1 Data Sets, Preprocessing, and Semi-Supervised Label Spreading

We employed a labeled dataset of CBC reports (1,281 reports, 14 standard CBC parameters, 9 diagnostic classes) and an unlabeled set of 643 regular CBC reports. The labeled data was split into 80% training and 20% held-out test subset. Diagnostic labels were encoded with `LabelEncoder`.

1.4.2 Ensemble Classifier and Hyperparameter Tuning

We did classification with a stacked ensemble of six base learners (RF, ET, SVC, LR, GB, and XGBoost) alongside a logistic regression meta-learner. We transmitted both the base-model probabilities and the engineering features to the meta-learner with passthrough=True. We optimized XGBoost using 5-fold RandomizedSearchCV and mitigated class imbalance with SMOTE–Tomek on the training dataset. To reduce variation, we executed the complete stacking technique thrice using random seeds (42, 202, and 777) and computed the mean predicted class probabilities for the test set.

1.4.3 Evaluation Metrics

The model has been evaluated on the held out test set, which was not used during pseudo-labeling, hyperparameter tuning, resampling, model fitting, or fitting of preprocessing. The discrimination performance was measured using accuracy, balanced accuracy, Cohen's kappa (κ), Matthews correlation coefficient (MCC), top-k accuracy for $k = 2, 3$, per-class precision, recall, F1-score, one-vs-rest ROC-AUC and PR-AUC. We evaluate calibration using multiclass log loss, Brier score, predicted calibration error (ECE), and reliability diagrams. We employed a confusion matrix with row-

Table 1.1 Per-class performance on the held-out test set.

Class	Prec.	Rec.	F1	Supp.
Healthy	0.98	0.97	0.98	67
Iron-def. anemia	0.95	0.97	0.96	38
Leukemia	0.88	0.78	0.82	9
Leukemia w/ thrombocytopenia	1.00	1.00	1.00	2
Macrocytic anemia	0.67	1.00	0.80	4
Normocytic hypochromic	0.96	0.95	0.96	56
Normocytic normochromic	0.95	0.98	0.96	54
Other microcytic anemia	0.82	0.75	0.78	12
Thrombocytopenia	1.00	0.93	0.97	15

normalization to observe error patterns class-wise. We used meta-learner coefficients, permutation importance of the final model, SHAP analysis, reliability testing and seed-level stability analysis to assess model explanations.

1.4.4 Quantitative Results

Overall and Per-Class Performance

The seed-averaged stacked ensemble obtained 94.94% accuracy on the held-out test set. The model also obtained balanced accuracy, Cohen's κ, MCC, and multiclass log loss of 0.9259, 0.9377, 0.9378, and 0.4097, respectively. The macro ROC–AUC was 0.9869 and the weighted ROC–AUC was 0.9826, which indicated good overall discrimination ability across the nine diagnostic classes. Table 1.1 shows the metrics for each class.

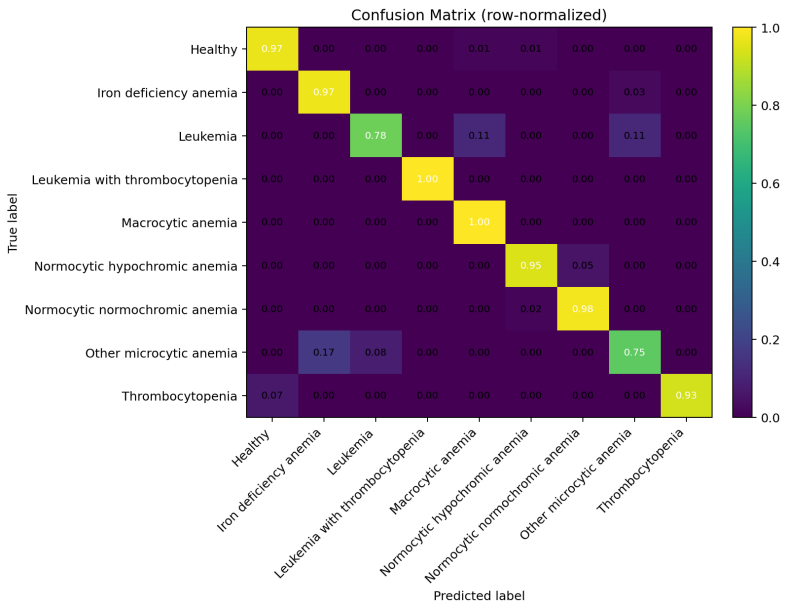


Figure 1.2 Row-normalized confusion matrix for the revised seed-averaged stacked ensemble.

Confusion Matrix and Error Analysis The row-normalized confusion matrix is shown in Figure 1.2. Healthy, iron deficiency anemia, leukemia with thrombocytopenia, macrocytic anemia, normocytic classes and thrombocytopenia were generally effectively classified. The main inaccuracies were in clinically overlapping classes. Normocytic hypochromic anemia was sometimes confused with normocytic normochromic anemia and other microcytic anemias were often confused with iron-deficiency anemia. Errors in leukemia were minimal but mostly against macrocytic anemia and other microcytic anemias. These patterns indicate that the other errors are not randomly distributed but are clustered in groups with similar CBC profiles.

Precision–Recall Curves and ROC The one-vs-rest ROC curves in Fig. 1.3 show strong discriminative performance across classes. Most classes achieved AUC values close to 1.00, with slightly lower but still strong AUC values for iron-deficiency anemia, leukemia, and normocytic hypochromic anemia. The precision–recall curves in Fig. 1.4 further show that most classes had high average precision. The lowest average precision values were observed for leukemia and other microcytic anemias, again indicating that these classes remain the most difficult to separate using CBC-only features.

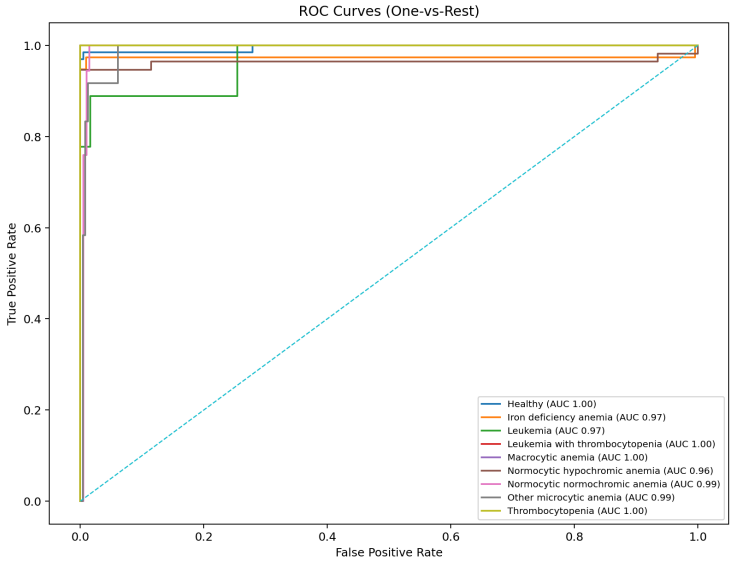


Figure 1.3 One-vs-rest ROC curves for all classes, demonstrating high discriminative capability of the proposed model across multiple hematological conditions.

Reliability and Calibration The model had reasonable probability calibration shown by = 0.4097 multiclass log loss and = 0.041 15-bin ECE. The reliability diagram (Fig. 1.5) reveals that the projected confidence was generally consistent with the empirical accuracy, although some bins were less stable due to the small number of test samples in the minority classes.

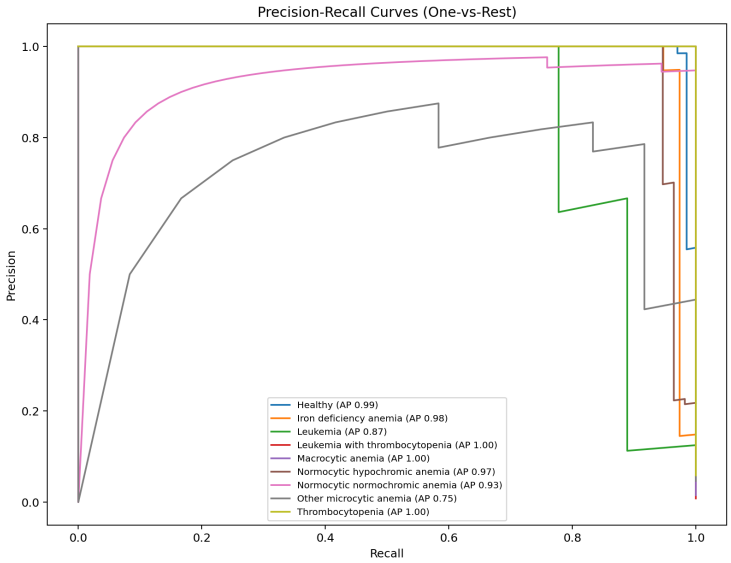


Figure 1.4 Precision–Recall curves for each class, highlighting performance under class imbalance and the model's ability to detect minority classes.

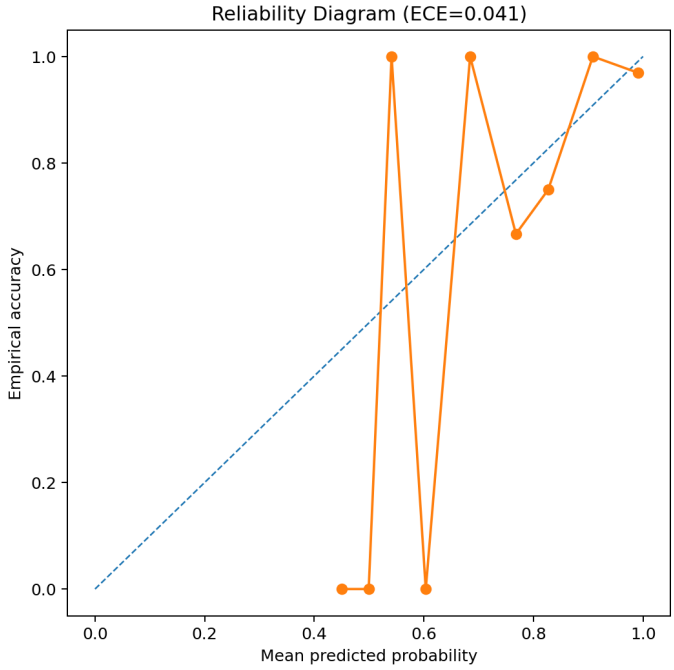


Figure 1.5 Reliability diagram

Interpretability and Feature Importance

The final model permutation importance showed that `MCV_micro_flag`, HGB, PLT, WBC, MCH, Shine–Lal index, HCT, MCHC, `LYM_abs`, and `HGB_low_flag` were

the most influential predictors. These features have clinical importance as they are markers of microcytosis, severity of anemia, platelet abnormalities, variation in leukocytes and red-cell morphology.

The SHAP analysis of the trained XGBoost component in the stack also showed clinically significant features. The top SHAP-ranked variables were HGB, PLT, `MCV_micro_flag`, MCV, WBC, `MCV_macro_flag`, `NEUT_abs`, MCHC, HCT and RBC. The concordance between permutation significance and SHAP supports the clinical credibility of the discovered decision patterns.

Stability and The reliability of XAI We also assessed the faithfulness and stability of the XAI results throughout several training runs. The final-model permutation importance identified `MCV_micro_flag`, HGB, PLT, WBC, MCH, Shine–Lal, HCT, MCHC, `LYM_abs`, and `HGB_low_flag` as the most influential predictors (see Fig. 1.6(A)). These are clinically significant traits, reflecting microcytosis, severity of anemia, platelet fluctuation, alterations in leukocytes and red-cell morphology.

We studied faithfulness by perturbing the top-k essential features and comparing the increase in log-loss to a random-k perturbation baseline. The removal of the top-ranked features consistently resulted in a higher performance reduction than the removal of random features (Fig. 1.6(B)), demonstrating that the detected features were indeed crucial for prediction. The stability was measured across the three stacked seed-specific models. Fig. 1.6(C) exhibits very high pairwise Spearman correlations, indicating the stability of explanatory patterns across seeds.

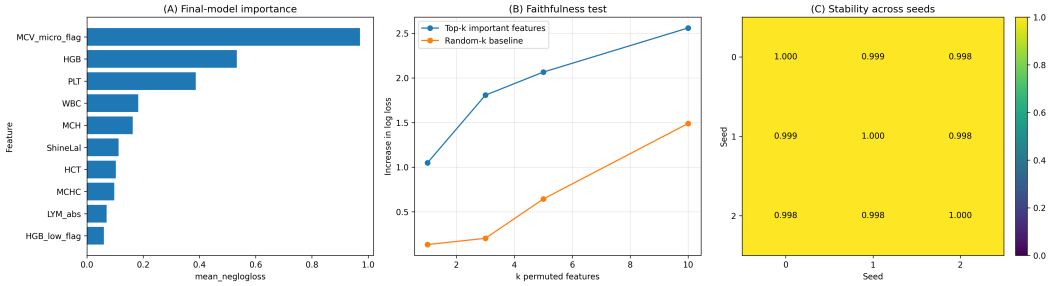


Figure 1.6 XAI evaluation of the revised final model: (A) final-model permutation importance, (B) deletion-based faithfulness test, and (C) seed-level stability of feature-importance rankings.

Class–specific SHAP analysis revealed that clinically significant CBC patterns predicted the distinct diagnoses. Macrocytic anemia was mainly connected to `MCV_macro_flag`, thrombocytopenia to PLT, and microcytic anemia-related classes to `MCV_micro_flag`, HGB, MCV, MCHC and related indices as shown in Fig. 1.7.

1.4.5 Comparative Analysis

Our methodology compares favorably to current CBC-based machine learning studies while dealing with a more difficult nine-class diagnostic assignment. In many previous works, great accuracy is reported for binary or limited multi-class issues, such as

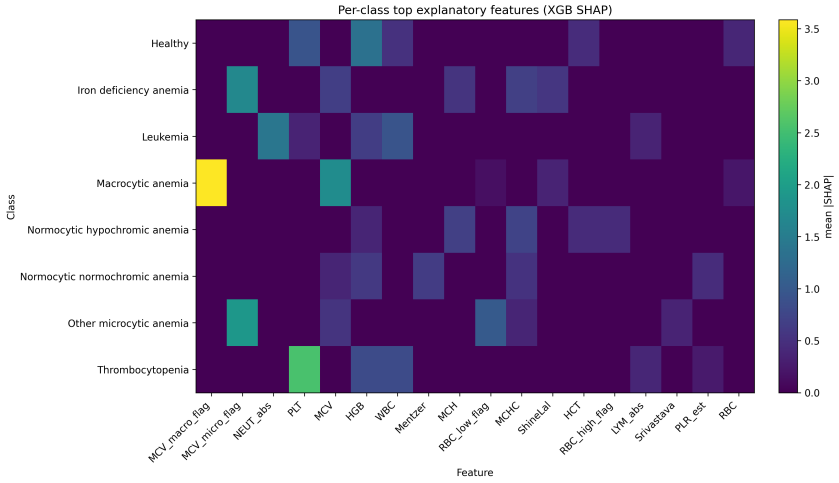


Figure 1.7 Per-class SHAP explanatory feature heatmap for the XGBoost component within the stacked ensemble.

anemia diagnosis, leukemia screening, or hemoglobinopathy prediction [17, 8]. Thus, a direct comparison is difficult due to different datasets, class definitions and validation settings.

Our final seed-averaged stacked ensemble obtained 94.94% accuracy and macro ROC–AUC of 0.9869 utilizing only routine CBC-derived features. This demonstrates that the pipeline can be simpler and more reproducible while reaching good performance with clinically designed features, semi-supervised pseudo-labeling and stacked ensemble learning.

The model strikes a pragmatic balance between accuracy, interpretability, calibration evaluation and methodological simplicity, though external validation would be necessary for clinical implementation.

1.4.6 Ablation Study

The proposed framework is constructed using optimized CBC features, high-confidence pseudo-labeled samples, SMOTE-Tomek balancing, and seed-averaged stacked ensemble learning.

Our model reached an accuracy of 94.94%, a balanced accuracy of 0.9259, an MCC of 0.9378, and a macro ROC–AUC of 0.9869 on the held-out test set. The XAI analysis also indicated stable and reliable feature significance patterns across random seeds. These results demonstrate that this pipeline is more straightforward, more interpretable, and more reproducible, while retaining high prediction performance.

1.5 CONCLUSION AND FUTURE WORK

In this study, a simplified CBC-based framework for multi-class hematological disease classification employing semi-supervised label spreading (SSL), clinically designed features and seed-averaged stacked ensemble was provided. The improved model

scored 94.94% accuracy, 0.9259 balanced accuracy, 0.9378 MCC, and 0.9869 macro ROC–AUC on the held-out test set.

The XAI study demonstrated that the model used clinically significant CBC markers such as `MCV_micro_flag`, HGB, PLT, WBC, MCH, MCV, and associated red-cell indices. Further analyses of steadfastness and seed stability validated the dependability of the explanation results. Overall, the framework achieves a superior mix of performance, interpretability and methodological simplicity.

Future studies will focus on external validation in independent institutions, analyzers and patient groups. Future developments may include prospective clinical evaluation, enhanced pseudo-labeling procedures and the use of additional hematological parameters if available.

ACKNOWLEDGEMENTS

We thank Mohaiminul Islam for guidance, the Consortium Research Center for funding and Sylhet MAG Osmani Medical College for access to unlabeled CBC data used in this study.

Bibliography

[1] E. Brihi, Normal and abnormal complete blood count with differential, 2024. Available: https://www.ncbi.nlm.nih.gov/books/NBK604207/

[2] J. Liu, Y. Gou, W. Yang, H. Wang, J. Zhang, S. Wu, S. Liu, T. Tao, Y. Tang, C. Yang, Development and application of machine learning models for hematological disease diagnosis using routine laboratory parameters: A user-friendly diagnostic platform, Frontiers in Medicine, vol. 12, p. 1605868, 2025. DOI: https://doi.org/10.3389/fmed.2025.1605868

[3] G. Airlangga, Leveraging machine learning for accurate anemia diagnosis using complete blood count data, Indonesian Journal of Artificial Intelligence and Data Mining, vol. 7, no. 2, pp. 318–326, 2024. DOI: http://dx.doi.org/10.24014/ijaidm.v7i2.29869

[4] A. Kumar, A. Srinath, G. Bharath, A. S. Chandra, R. S. Rani, M. Manikandan, S. Mohandoss, Data to diagnosis machine learning models for accurate anemia classification, International Journal of Scientific Research and Technology, 2025. DOI: https://doi.org/10.5281/zenodo.15204106

[5] S. Pullakhandam and S. McCoy, Classification and explanation of iron deficiency anemia from complete blood count data using machine learning, BioMedInformatics, vol. 4, no. 1, pp. 661–672, 2024. DOI: https://doi.org/10.3390/biomedinformatics4010036

[6] J. Lv, J. Li, X. Ren, Q. Huang, S. Deng, Machine learning for discriminating microcytic hypochromic anemia based on erythrocyte parameters, International Journal of Laboratory Hematology, 2025. DOI: https://doi.org/10.1111/ijlh.14524

[7] W. Tepakhan, W. Srisintorn, T. Penglong, P. Saelue, Machine learning approach for differentiating iron deficiency anemia and thalassemia using random forest and gradient boosting algorithms, Scientific Reports, vol. 15, no. 1, p. 16917, 2025. DOI: https://doi.org/10.1038/s41598-025-01458-5

[8] A. Schipper, M. Rutten, A. van Gammeren, C. L. Harteveld, E. Urrechaga, F. Weerkamp, G. den Besten, J. Krabbe, J. Slomp, L. Schoonen, Machine learning-based prediction of hemoglobinopathies using complete blood count data, Clinical Chemistry, vol. 70, no. 8, pp. 1064–1075, 2024. DOI: https://doi.org/10.1093/clinchem/hvae081

[9] Z. J. Cheng, H. Li, M. Liu, X. Fu, L. Liu, Z. Liang, H. Gan, B. Sun, Artificial intelligence reveals the predictions of hematological indexes in children with acute leukemia, BMC Cancer, vol. 24, no. 1, p. 993, 2024. DOI: https://doi.org/10.1186/s12885-024-12646-3

[10] F. Christensen, D. K. Kilic, I. E. Nielsen, T. C. El-Galaly, A. Glenthoj, J. Helby, H. Frederiksen, S. Moller, A. D. Fuglkjaer, Classification of alpha-thalassemia data using machine learning models, Computer Methods and Programs in Biomedicine, vol. 260, p. 108581, 2025. DOI: https://doi.org/10.1016/j.cmpb.2024.108581

[11] R. Z. Haider, I. U. Ujjan, N. A. Khan, E. Urrechaga, T. S. Shamsi, Beyond the in-practice CBC: the research CBC parameters-driven machine learning predictive modeling for early differentiation among leukemias, Diagnostics, vol. 12, no. 1, p. 138, 2022.

[12] Y. Cakmak, I. Pacal, AI-driven classification of anemia and blood disorders using machine learning models, Computers and Electronics in Medicine, vol. 2, no. 2, pp. 43–52, 2025. DOI: https://doi.org/10.69882/adba.cem.2025073

[13] Y. Cakmak, Machine learning approaches for enhanced diagnosis of hematological disorders, Computational Systems and Artificial Intelligence, vol. 1, no. 1, pp. 8–14, 2025. DOI: https://doi.org/10.69882/adba.csai.2025072

[14] R. J. Hindi, Ensemble machine learning approach for anemia classification using complete blood count data, Al-Mustansiriyah Journal of Science, vol. 36, no. 3, pp. 51–70, 2025. DOI: https://doi.org/10.23851/mjs.v36i3.1709

[15] J. Prajapati, V. Uduthalapally, D. Das, R. Mahapatra, P. N. Wasnik, XAIA: An explainable AI approach for classification and analysis of blood anemia, in 2023 OITS International Conference on Information Technology (OCIT), pp. 88–93, 2023. DOI: https://doi.org/10.1109/OCIT59427.2023.10430938

[16] A. S. Awaad, Y. M. Elbarawy, H. Mancy, N. E. Ghannam, Exploring CBC data for anemia diagnosis: A machine learning and ontology perspective, BioMedInformatics, vol. 5, no. 3, p. 35, 2025. DOI: https://doi.org/10.3390/biomedinformatics5030035

[17] H. Amjad, Z. Hussain, M. Hasan, M. Ul Hassan, Machine learning-based models for screening of anemia and leukemia using features of complete blood count reports, Scientific Reports, vol. 15, no. 1, p. 33333, 2025. DOI: https://doi.org/10.1038/s41598-025-21279-w

[18] M. U. Nasir, M. Zubair, M. T. Naseem, T. Shahzad, A. Saeed, K. M. Adnan, A. H. Gandomi, Multiclass classification of thalassemia types using complete blood count and HPLC data with machine learning, Scientific Reports, vol. 15, no. 1, p. 26379, 2025. DOI: https://doi.org/10.1038/s41598-025-06594-6

[19] I. Uvaliyeva, Z. Rakhmetullina, D. Borozenets, F. Amenova, S. Tezekpayeva, Application and evaluation of ensemble machine learning models for the differential diagnosis of clinical and hematological syndromes in the health passport software package, Engineered Science, vol. 36, p. 1635, 2025. DOI: http://dx.doi.org/10.30919/es1635

[20] R. Zahorec, Ratio of neutrophil to lymphocyte counts—rapid and simple parameter of systemic inflammation and stress in critically ill, Bratislavske Lekarske Listy, vol. 102, no. 1, pp. 5–14, 2001. Available: https://pubmed.ncbi.nlm.nih.gov/11723675/

[21] A. Y. Gasparyan, L. Ayvazyan, U. Mukanova, M. Yessirkepov, G. D. Kitas, The platelet-to-lymphocyte ratio as an inflammatory marker in rheumatic diseases, Annals of Laboratory Medicine, vol. 39, no. 4, pp. 345–357, 2019. DOI: https://doi.org/10.3343/alm.2019.39.4.345

[22] J. M. England, P. M. Fraser, Differentiation of iron deficiency from thalassaemia trait, The Lancet, vol. 301, no. 7818, p. 1514, 1973. DOI: https://doi.org/10.1016/s0140-6736(73)91856-4

[23] I. Shine, S. Lal, A strategy to detect beta-thalassaemia minor, The Lancet, 1977. DOI: https://doi.org/10.1016/s0140-6736(77)92128-6

[24] R. Green, R. King, A new red cell discriminant incorporating volume dispersion for differentiating iron deficiency anemia from thalassemia minor, Blood Cells, vol. 15, no. 3, pp. 481–491, 1989. Available: https://pubmed.ncbi.nlm.nih.gov/2620095/

[25] T.-C. Huang, Y.-Y. Wu, Y.-G. Chen, S.-W. Lai, C.-S. Liu, Discrimination index of microcytic anemia in young soldiers: A single institutional analysis, PLOS ONE, vol. 10, no. 2, p. e0114061, 2015. DOI: https://doi.org/10.1371/journal.pone.0114061

[26] B. Hu, X.-R. Yang, Y. Xu, Y.-F. Sun, C. Sun, W. Guo, X. Zhang, Systemic immune-inflammation index predicts prognosis of patients after curative resection for hepatocellular carcinoma, Clinical Cancer Research, vol. 20, no. 23, pp. 6212–6222, 2014. DOI: https://doi.org/10.1158/1078-0432.ccr-14-0442