Al-Mustansiriyah Journal of Science
ISSN: 1814-635X (print) 2521-3520 (electronic)

ORIGINAL ARTICLE OPEN ACCESS

Ensemble Machine Learning Approach for Anemia Classification Using Complete Blood Count Data

Rasha Jamal Hindi

Computer Science, College of Education, Mustansiriyah University, Baghdad, Iraq

CORRESPONDANCE
Rasha Jamal Hindi
rashajamal94@uomustansiriyah.edu.iq

ARTICLE INFO
Received: Jul. 15, 2025
Revised: Sep. 16, 2025
Accepted: Sep. 20, 2025
Published: Sep. 30, 2025

ABSTRACT: ***Background:*** *Anemia is a widespread global health issue that affects millions of individuals worldwide. Early and accurate diagnosis is critical for effective treatment. Traditional diagnostic approaches rely on complete blood count (CBC) parameters, which provide valuable clinical insights but may require advanced tools to enhance diagnostic accuracy.* ***Objective:*** *This study aims to develop and evaluate machine learning models for the classification of different anemia subtypes using CBC data. The objective is to assess the performance of individual models and ensemble methods in improving diagnostic accuracy.* ***Methods:*** *Five machine learning algorithms were implemented for the classification task: Decision tree, random forest, XGBoost, gradient boosting, and neural networks. In addition to individual model evaluation, ensemble techniques including hard voting, soft voting, and stacking were applied to enhance model performance.* ***Results:*** *Experimental results showed that ensemble methods significantly outperformed individual models in classification accuracy. Among them, the stacking ensemble achieved the highest accuracy of 98.44%, indicating superior performance in distinguishing anemia subtypes.* ***Conclusions:*** *The study demonstrates that ensemble learning methods, particularly stacking, can substantially improve the performance of machine learning models in anemia classification based on CBC data. These findings suggest the potential integration of such ensemble techniques into clinical decision-support systems to assist healthcare providers in making efficient and timely diagnoses.*

KEYWORDS: *Anemia classification; Machine learning; Complete blood count; Ensemble methods; Decision tree*

INTRODUCTION

Anemia is one of the most common diseases in the world [1]–[5], which together with the correct and timely diagnosis is important for the further management of the disease among patients of different ages. Anemia is known to be a decline in the number of red blood cells or their ability to transport oxygen throughout the body. This is not a condition that merely affects the lives of single people but also has repercussions on the healthcare facilities globally. It becomes harder to identify anemia given the many factors that lead to its development, such as deficiency of iron, genetic disorders, infections, chronic diseases among others. Conventional diagnosis of anemia involves complete blood count (CBC) tests [6]–[8], which are key hematological markers including the hemoglobin (HGB) level, red blood cell count (RBC), and mean corpuscular volume (MCV). However, the manual evaluation of these parameters is often time-consuming and subject to errors, especially when it comes to the analysis of the different types of anemia. The utilization of machine learning (ML) as an instrument for augmenting diagnostic performance has attracted interest as a result of the inadequacy of current diagnostic methods. Intelligent computation techniques such as ML can identify complex patterns of data in CBC and hence provide faster and more accurate differentiation of anemia subtypes. In this paper, the use of ML models specific to predicting and categorizing anemia types using CBC data is discussed. We compare decision tree, Random Forest, XGBoost, gradient boosting, and Multi-Layer Perceptron (MLP) algorithms to address the anemia classification situation. Moreover, different types of ensemble learning techniques like hard voting, soft voting, and stacking are used

to improve the model accuracy [9], [10]. Combined methods take advantage of the use of individual models [11]. It proves that ensemble methods in turn provide improved generalization, especially in cases when an increase in class imbalance is observed, which is characteristic of most medical datasets [12].

The goal of this work is to find a consistent anemia classification and a fully automated diagnostic system that can be helpful for healthcare professionals. It also aims to enhance the diagnostic accuracy by applying the ensemble learning approaches, as well as to overcome the main issues, class imbalance problem and interpretation of the models. This study makes the following original contributions compared to existing literature: Reliance exclusively on CBC data for anemia subtype classification, systematic comparison of a wide range of ML models under a unified framework, introduction of multiple ensemble strategies including stacking, prioritization of minority class detection through class imbalance handling, and incorporation of local interpretable model-agnostic explanations (LIME)-based explainability to ensure clinical interpretability.

The novelty of this study lies in the systematic integration of ensemble strategies specifically tailored to anemia subtype classification from CBC data. Unlike previous works that typically applied single classifiers or basic ensemble voting, our configuration combines decision tree, Random Forest, and XGBoost as base learners with Logistic Regression as a meta-learner in a stacking framework. This design was chosen to leverage the complementary strengths of tree-based models interpretability, robustness, and handling of non-linear feature interactions while allowing the meta-learner to correct for their residual errors. This contribution demonstrates that carefully configured ensemble methods can provide a clinically relevant advancement in automated anemia diagnosis. The structure of this paper is as follows: Section 3 sheds light on the research work done in anemia prediction, along with the use of ML technologies in medical diagnosis. The paper focuses on the method section and devotes Section 4 to the data collection, data pre-processing, and model building. The results and discussion in Section 5 give a detailed analysis of the models' performance, where confusion matrices and classification reports were produced. The conclusion and future work are presented to reveal potential improvements for anemia diagnosis with the help of ML.

RELATED WORKS

This section surveys existing research works in anemia prediction and its sub-domains, including ML-based solutions for improving diagnostic accuracy, as well as limitations of anemia classification problems and an ensemble learning approach to medical diagnostics. HGB levels below the specified reference level for age and sex were strictly the definition of anemia [13]–[16]. Anemia is a major public health issue affecting over 1.98 billion individuals globally [17]. Iron deficiency is the most common cause of anemia, resulting in decreased production of red blood cells [18]. Iron supplementation is common practice, although diagnosis of anemia by regular blood tests (including CBC) is still practical [19]. Apparently, many studies have investigated the ability of ML algorithms in identifying anemia. Previous works used decision trees, random forests, and support vector machines (SVMs) for classifying anemia in blood data [20], [21]. While these models performed well in predicting the presence of anemia, they were less accurate for distinguishing iron deficiency anemia (IDA) from other etiologies of anemia, such as genetic disorders [22]. In the last few years, research has been carried out with neural network models and feature selection methods, such as ensemble learning techniques, to refine anemia subtype classification [23].

In addition, several ML models have been used to predict types of anemia from CBC data alone without relying on the more expensive serum ferritin tests [24], [25]. As an example, recent studies have effectively implemented random forest models to predict whether or not patients are low in ferritin with high sensitivity and specificity in those diagnosed with IDA. Validation of these models in larger datasets is needed to more thoroughly prove that their generalizability holds across various age and gender subgroups [26]. The authors have taken steps forward based on these by implementing different ML models and ensemble techniques to enhance the prediction of anemia using CBC data only. Moreover, ensemble learning has been recently used for medical diagnostics to aggregate multiple models and make a decision based on the opinion of the majority of learners for higher accuracy [27]. Ensemble learning has been adapted to disease diagnosis problems as well [28], [29].

For instance, a study by Rane *et al.* developed a voting classifier model able to discriminate β-thalassemia carriers from control individuals using red blood cell indices, obtaining an accuracy of 93%. Similarly, other studies have shown the accuracy of ensemble methods in thalassemia and blood disorder detection [27] as well as using deep learning frameworks. Ensemble techniques have been reported to be more accurate than individual classifiers within different medical domains, as demonstrated by studies focused on anemia prediction. For instance, Saleem *et al.* used an ensemble

of classifiers including K-nearest neighbors, decision trees, and gradient boosting, which performed better for thalassemia [30]. This paper novelly introduces different strategies of ensembles (hard-voting, soft-voting, and stacking) into anemia classification tasks using varying base CNN models, illustrating a superior ability of the ensemble models to enhance diagnostic accuracy compared to individual classifiers in facilitating diagnosis, especially for minority classes (such as rare types of anemia). Class imbalance problem in anemia prediction, for example, rare types like Leukemia with Thrombocytopenia have much less samples than some anemia types (e.g. IDA) because of which there will be a high class imbalance in the data, inevitably leading to poor model performance [31]. Class imbalance, which can cause a model to overfit on the majority classes at the expense of the minority classes, decreasing diagnostic accuracy overall [32]. To counter these problems, class balancing techniques have been used in studies such as over-sampling, under-sampling, and synthetic data generation to improve the model performance for all classes [27], [33]. Authors of [27], [33] have also prioritized precision, recall, and F1 score to make sure the models are not only accurate but have high sensitivity and specificity, especially for minority classes. An emerging trend in medical treatment we have on our hands is the use of explainable AI (XAI) models for interpretable ML. As noted by Alharthi *et al.*, it is particularly important to have interpretable models in a clinical setting where knowledge of the reasons behind what leads to a diagnosis can play an essential role in good decision-making [34]. Table 1 presents an overview of related works.

Table 1. Overview of related works

Study	Focus	ML Techniques	Results/Findings	Challenges Addressed
Rane *et al.* [27]	Classification of β-thalassemia carriers	Ensemble learning (Voting: SVM, gradient boosting, Random Forest)	Achieved 93% accuracy in detecting β-thalassemia carriers using red blood cell indices	Improved diagnostic accuracy through ensemble methods
Saleem *et al.* [30]	Thalassemia prediction	Combination of K-Nearest Neighbors, decision trees, gradient boosting	Higher predictive accuracy for thalassemia compared to individual classifiers	Enhanced performance on minority classes using ensemble models
Alharthi *et al.* [34]	Explainable AI in medical diagnostics	SAELM Hybrid Algorithm	Improved prediction of thalassemia	Application of explainable models for clinical decision-making
Nair *et al.* [28]	Thalassemia detection using non-invasive methods	Machine learning with optoelectronic measurements	Non-invasive detection approach, eliminating the need for blood tests	Non-invasive and innovative detection method
Laeli *et al.* [35]	Hyperparameter optimization in SVM for thalassemia classification	Grid search optimization for SVM	Enhanced classification accuracy through hyperparameter tuning	Optimal performance of SVM for thalassemia detection
Abdulkarim *et al.* [29]	Prediction of thalassemia	Hybrid algorithm combining SAELM	Improved classification accuracy	Overcoming model limitations by combining learning techniques
Dugusheva *et al.* [24]	Anemia diagnosis	Diagnostic tests, including CBC parameters	Focus on improving non-invasive diagnostic methods for anemia	Diagnostic tests for specific anemia conditions

MATERIALS AND METHODS

The procedure used in this study is described in Figure 1, showing the general procedure for anemia prediction using the ML technique. The process starts with the data loading as well as the exploratory data analysis (EDA) phase, during which the information about the given dataset is presented, including the class distribution and missing values. Following EDA, data transformation steps such as data scaling or normalization, feature extraction, and rebalancing of classes are used for a dataset before feeding it into the learning model. Subsequently, seven major ML models, which are SVM, extra

trees, random forest, decision tree, XGBoost, gradient boosting, and MLP are invoked to be trained with the preprocessed data. In order to consider these models, main factors like time taken to train, accuracy, precision measures like F1 score, error percentage, and recall are taken into consideration. The last process in the methodology is the process of taking the results of the models with ensemble learning as a final step. More enhancements of the voting technique include hard voting, soft voting, and stacking ensembles to enhance the best-performing classifiers while avoiding their drawbacks.

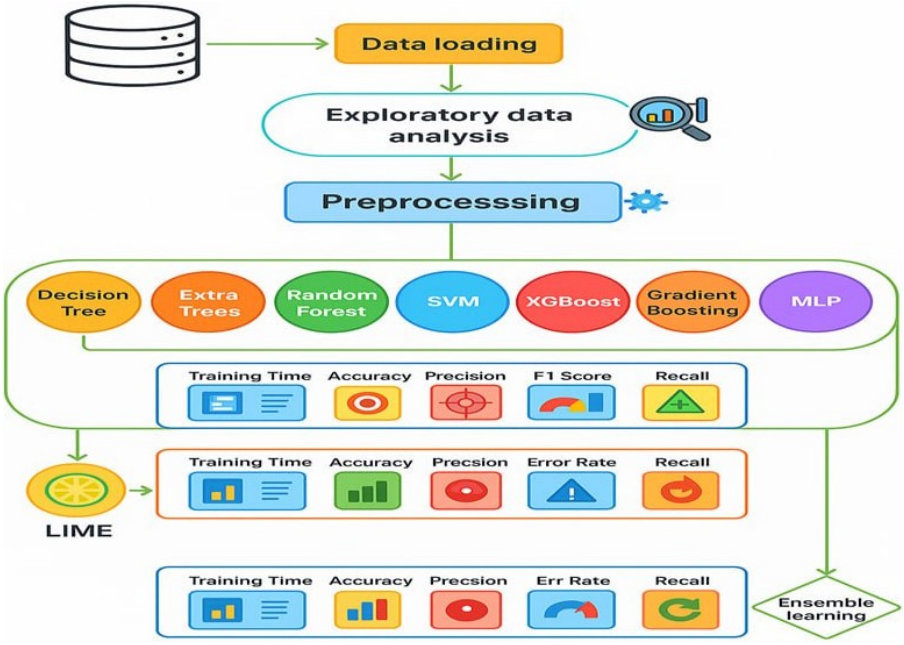

Figure 1. Proposed methodology

Data Acquisition

When choosing the data in a study, which is usually a homework assignment in ML research, especially in the medical area, it has to be of high quality. For this study, the anemia types classification dataset from Kaggle was considered, which includes CBC data that is commonly used to diagnose the different types of anemia. The dataset consists of a feature vector of CBC and the anemia type they belong to, namely Iron-deficiency anemia, aplastic anemia, and hemolytic anemia, with RBC, HGB, hematocrit (HCT), and other blood parameters. Consistent with this, the dataset includes 14 input features that are rate-independent: WBC, LYMp, PLT, etc., and 1 rate target variable, diagnosis, classifying the type of anemia. The collected data can be used for academic and commercial purposes since it is licensed under Apache-2.0. Observed from clinical prescriptions, the dataset shows authentic hematological data and therefore, can be used to build predictive models to distinguish between different types of anemia. The data was then retrieved effectively using the Kaggle API and copied to a local environment for analysis. The first exploratory analyses verified the presence of missing values, data correctness, and class distribution, demonstrating that some classes are unbalanced. These were addressed during the preprocessing phase when we were granted permission to balance the models and prevent biased results. The anemia types classification dataset was obtained from Kaggle, containing CBC data for various anemia subtypes. The dataset includes 14 input features (e.g., RBC, HGB, HCT, PLT, WBC) and one categorical target variable (Diagnosis).

Exploratory Data Analysis (EDA)

EDA [36]–[38] is the process of attaining an initial understanding of the nature of the data before any modeling. EDA was done to establish the distribution of cases of anemia diagnosis, and realized that the classes are balanced and there are no missing values in the selected dataset. Another decision that was made possible by EDA was to get to know the distribution and frequency of each of the anemia types, which can influence decisions on data preprocessing and model selection.

Class Distribution

In the bar plot and pie chart as presented in Figure 2, there is a clear demonstration that the classes are not evenly distributed in the various types of anemia. The largest class in the dataset is the Healthy class, including 26.2% of the cases, and ranked sequentially by the normocytic hypochromic anemia (NHA), including 21.8% of the cases, and then the normocytic normochromic anemia that includes 21.0% of the cases [39]. Combined with these three classes, they represent a majority of the data, which indicates class imbalance where certain rare anemias, such as Leukemia with Thrombocytopenia and macrocytic anemia, have less than 2% of the samples.

This imbalance is also evident in the bar chart, with the Healthy class having over 300 instances and the two samples in Leukemia and Other Microcytic anemia having even lesser samples with Leukemia with Thrombocytopenia having only 11 samples. This matters most when corrected prior to model training, with models possibly preferring the majority classes in prediction, as they would offer poor accuracy for the minority classes. Such pre-processing methods like down-sampling, oversampling, or class weighting might be necessary in order to help the models learn well across the different types of anemia. EDA was performed to assess data quality and distribution before model training. The dataset consisted of 14 CBC features and a categorical target variable indicating anemia type. The analysis revealed no missing values, confirming the completeness of the dataset. Class distribution analysis, visualized through bar and pie charts, highlighted a strong imbalance: Common categories such as Healthy (26.2%), NHA (21.8%), and Normocytic Normochromic anemia (21.0%) represented the majority, while rare subtypes such as Leukemia with Thrombocytopenia (<2%) and Macrocytic anemia were severely underrepresented. This imbalance was addressed in preprocessing through resampling techniques. Overall, EDA confirmed that the dataset is clean and reliable, but emphasized the need for balancing strategies to ensure fair model performance across both common and rare anemia classes.

To decipher the various associations of different anemias to show whether they deviate from the norm or exhibit other trends, figures including bar graphs and pie graphs were employed (as illustrated in Figure 2). The bar chart clearly indicates the fact that the frequency distribution of the diagnosis type is positively skewed, and hence the existence of large variation in the frequency of each diagnosis type. This visualization was useful for identifying those specific classes that would need different treatment during the preprocessing stage.

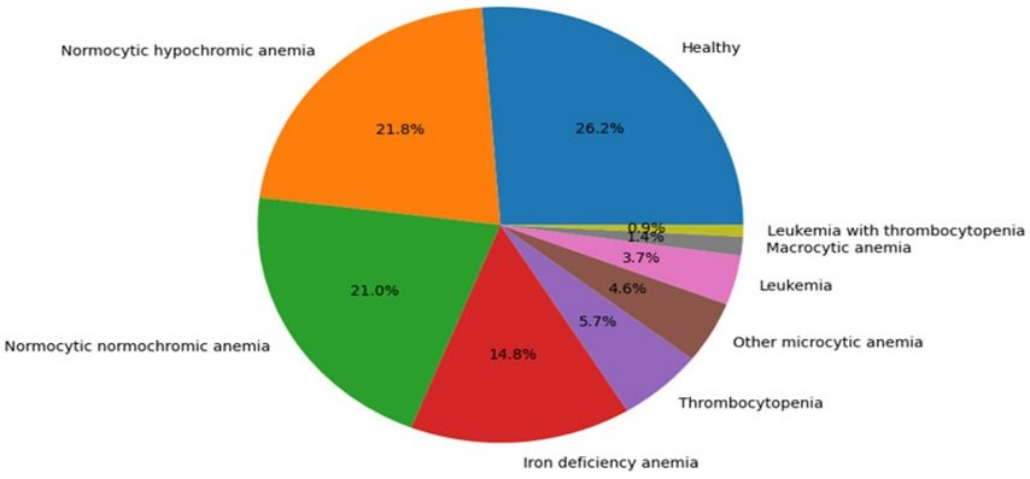

Figure 2. Visualization of different classes

In fact, a pie chart is a perfect illustration of the relative proportion of the overall dataset and the proportion in each diagnostic type. For example, compared with other PA types, it points to the prevalence of some common types, such as Healthy, NHA, and IDA, while providing a graphic indication of the minority status of Macrocytic anemia and Leukemia with Thrombocytopenia.

Missing Values Analysis

An essential aspect of EDA is identifying missing or incomplete data. The dataset was checked for missing values across all columns. Thankfully, the dataset contained no missing values, which

eliminated the need for imputation techniques that could potentially introduce noise into the data. The lack of missing data ensured that all features could be used as inputs for the ML models without any additional data handling or cleaning steps. This consistency across the dataset strengthens the robustness of the analysis, allowing the models to learn from the full scope of information available from the CBC test results.

Insights from EDA

EDA revealed several critical characteristics of the dataset. First, a pronounced class imbalance was identified, with the majority of instances belonging to frequently occurring categories such as Healthy and Normocytic Hypochromic Anemia. Addressing this imbalance is a necessary preprocessing step to mitigate potential model bias. Second, visualizations provided clear insights into the data structure: bar charts effectively illustrated the distribution of anemia types, while pie charts were instrumental in analyzing the compositional makeup of the dataset, thereby informing subsequent preprocessing and modeling strategies. Finally, the analysis confirmed the dataset's completeness, as no missing values were found across any features. This integrity ensures the data is clean and immediately suitable for model development, eliminating the need for extensive data cleaning procedures.

Data Preprocessing

Data preprocessing is an essential aspect of the ML model creation process [40], [41], even more so for medical applications, where data quality determines the accuracy of prognoses. In this work, several preprocessing techniques were performed on the anemia dataset before training ML models. Some of them include: Scaling of features, encoding of the target variable, and class imbalance, which are very important if the model is to perform optimally.

The independent variables in the dataset include different parameters of the CBC, and these parameters are measured on different scales. For example, WBC count can be in thousands while MCH is measured in picograms. This variation in feature scales can have a deleterious effect on model performance, particularly algorithms that depend on distance calculations, such as SVM or neural networks. In this regard, feature scaling was used to bring the range of values of all features into a standard range. Minmax was scaling here, where all the feature values were normalized so that they lay between 0 and 1. This normalization aids models to be trained faster and further restrains the features with huge scales to have an enormous impact in the learning phase.

Diagnosis is the target variable, which consists of the categorical values of anemia types. This was important because most ML algorithms can only accept numerical inputs, and therefore, the target variable needed to be transformed into a numerical form. In order to be used within the classification algorithms, label encoding was used to map each anemia type to a distinct numerical value. This encoding method did not degrade the class labels in a manner that introduced complication, so that the model could distinguish between the types of anemia it was supposed to differentiate.

One of the main challenges observed while exploring the dataset for the first time was related to an imbalance in classes of the target variable. Some types of anemia occurred more frequently compared to others, which might cause bias towards the predictions optimizable by a model (optimal flat models tend to predict majority classes mostly, and minority class predictions are often sacrificed). This issue was alleviated by balancing the composition of the dataset by class-balancing techniques. Random under-sampling is a popular strategy where the common instances of the majority class are reduced to balance it with minority classes in such a way that overall, you have a balanced dataset. An alternative to under-sampling is random oversampling, which randomly duplicates examples from the minority class to balance the class distributions. In clinical diagnosis tasks, it is important to balance the class distribution (to avoid very rare cases being disproportionately misclassified by a trained model, which could have catastrophic real-world implications, e.g., missing serious life-threatening forms of anemia).

After preprocessing the data, it was split into training as well as testing datasets. This split is required for testing the model on new, unseen data. We used an 80–20 split, meaning the data was broken into an 80% training set and a 20% testing dataset. It was stratified so as to ensure the distribution of classes within ranks remained similar for both training and testing sets, aiding in removing any bias during evaluation. This step guarantees that the trained models generalize well in new data, unknown by now, and do not learning a pattern in the training set. The stratified train test split also helps in keeping the minority classes of the data intact. The dataset exhibited strong class imbalance, with common categories such as "Healthy" (326 instances) and "NHA" (261 instances) dominating, while rare conditions such as "Leukemia with Thrombocytopenia" contained only 11 samples. To

mitigate this imbalance, we tested multiple balancing strategies. For tree-based models (random forest, Extra Trees, XGBoost, gradient boosting, decision tree), random oversampling of minority classes was applied to prevent bias towards majority classes. For SVM and MLP, which are more sensitive to duplicated samples, random under-sampling of the majority classes was performed to maintain training stability. We did not include synthetic sampling methods such as SMOTE or ADASYN in the current experiments, but we acknowledge their potential utility and list them as a direction for future work. Despite these measures, very rare classes (e.g., Leukemia with Thrombocytopenia) remained difficult to classify reliably, as reflected in low F1-scores, which highlights the inherent challenge of working with highly imbalanced medical data.

Data validation occurred after each stage to confirm that the preprocessing steps had not added any errors or inconsistencies. This consisted of assessing feature transformation integrity, correcting class imbalance, and checking whether the training and testing sets were representative of the full sample. This validation, in turn, helps guard against issues like data leakage, the test set failing to remain blind, and information from the test set being used directly or indirectly by the model to determine its parameters, thereby providing overly optimistic performance estimates.

Model Training

Model training is a crucial step in ML where the algorithm learns patterns from data and makes predictions, and it helps us train the model. We used these models to classify anemia on the CBC data into different types. All the selected models are tree-based classifiers [42]–[44], boosting methods [45]–[47], SVM [48]–[52], and neural networks [53]–[59], each of which has its own strength in terms of accuracy with stability, and computational efficiency.

Selected Models

A diverse set of machine learning models, representing a wide variety of algorithmic approaches, was selected for the anemia classification task. This selection ensures a comprehensive evaluation of different methodologies.

1 Tree-Based Models

The Decision Tree Classifier served as a foundational, interpretable model. It operates by recursively splitting the dataset based on feature importance, providing clear decision paths that are valuable in medical contexts where model explainability is crucial.

Ensemble methods built upon this foundation to enhance performance. The Random Forest Classifier constructs a multitude of decision trees and aggregates their predictions. This non-linear approach effectively handles high-dimensional feature spaces, making it well-suited for complex CBC data containing numerous parameters. The extra trees classifier (Extremely Randomized Trees) further promotes diversity among its trees by using random splits on features, a technique that reduces overfitting and often improves generalization to unseen data.

2 Gradient Boosting Models

Two gradient boosting implementations were employed. XGBoost is an optimized algorithm that builds trees sequentially, with each new tree correcting the errors of its predecessors. Its advanced regularization techniques make it highly effective for tasks with complex decision boundaries, such as anemia classification. The Gradient Boosting Classifier follows a similar sequential methodology but typically uses fewer subsampling techniques, resulting in a model that is less complex and less prone to overfitting, though often computationally slower.

3 Non-Tree-Based Models

To provide algorithmic diversity, a SVM was included. This algorithm finds an optimal hyperplane to maximally separate classes in a high-dimensional space, making it powerful for both linear and non-linear classification problems. Finally, a Multi-Layer Perceptron (MLP), a class of neural network, was used to capture highly complex, non-linear relationships within the CBC parameters through its multiple layers of interconnected neurons.

 DOI: https://doi.org/10.23851/mjs.v36i3.1709

Training Process

A structured pipeline was implemented to ensure a consistent, reliable, and reproducible evaluation of all models. Each model was initialized with a predefined set of hyperparameters, informed by established literature and preliminary experimentation. To guarantee reproducibility, models incorporating stochastic elements, such as Random Forest and Extra Trees, were initialized with a fixed random seed. For gradient boosting implementations (XGBoost and Gradient Boosting), critical parameters including learning rate, the number of estimators, and maximum tree depth were configured to optimize performance.

The preprocessed and shuffled dataset was partitioned into a training set (80%) and a held-out test set (20%). This split ensured sufficient data for model induction while reserving a representative subset for an unbiased evaluation of generalizability to unseen data. Model performance was rigorously assessed on the test set using a comprehensive suite of metrics: accuracy, precision, recall, F1 score (the harmonic mean of precision and recall), and error rate. These metrics provided a multifaceted view of predictive efficacy across different anemia types.

To mitigate overfitting and ensure robust generalization, k-fold cross-validation was employed during the model training phase. Furthermore, a systematic hyperparameter optimization was conducted. A grid search strategy was applied to tree-based models (Decision Tree, Random Forest, Extra Trees) over parameters such as maximum depth, minimum samples per split, and the number of estimators. For the more computationally complex boosting models (XGBoost, Gradient Boosting), a random search with 50 iterations was utilized to explore combinations of the learning rate, maximum depth, number of estimators, and subsampling ratio. The Support Vector Machine (SVM) was tuned via grid search over the kernel type, regularization parameter C, and kernel coefficient . The Multi-Layer Perceptron (MLP) was optimized using a random search across the number of hidden layers, neurons per layer, activation function, and learning rate. The optimal hyperparameter set for each model was selected based on the best average performance achieved through 5-fold cross-validation on the training data. These tuned models were subsequently evaluated on the independent test set for final performance comparison.

Evaluation and Model Performance

Model performance was rigorously evaluated on a held-out test set following the training phase. The evaluation employed a suite of metrics to provide a comprehensive assessment of predictive accuracy and reliability. A confusion matrix was generated for each model to provide a granular view of its classification behavior. This matrix facilitated the analysis of true positives (Tp), true negatives (Tn), false positives (Fp), and false negatives (Fn) for each anemia type, enabling a detailed examination of per-class performance. This analysis was complemented by a full classification report for every model, which detailed key metrics for each class: precision, recall, F1-score, and support. These reports were instrumental in evaluating the efficacy of each model across all classification tasks, with particular attention paid to the performance on minority classes to ensure the models did not solely favor the majority diagnoses.

The results for all models were consolidated into a comprehensive summary table to facilitate direct comparison. This table ranked the models based on overall accuracy and other key performance metrics, providing a clear hierarchy of their predictive capabilities. To augment this tabular data, the models were further compared through visualizations. An accuracy bar plot provided an immediate, clear comparison of overall model performance, allowing for the rapid identification of the most accurate algorithms for anemia prediction. Beyond aggregate accuracy, a detailed classification report—listing precision, recall, F1-score, and support for each class—was computed for every model. This report was instrumental in evaluating performance on a per-class basis, which was critical for assessing the models' efficacy in detecting the more challenging minority anemia classes and ensuring a robust and equitable diagnostic capability. The findings revealed that XGBoost was found to be the most robust model for anemia type prediction as it was consistently accurate, precise, and had with high recall compared to other models. It outperformed FR and LR because of the fact that it can deal with imbalanced datasets, and it was able to capture the complex pattern in CBC data. We found that random forest and extra trees were the other two high-performing models with significantly shorter training times, great candidates in scenarios where model interpretability and computational efficiency are important. Two models, the SVM and MLP, while providing decent results for both time of completion metrics, appear to be more computationally expensive than tree-based approaches, where further hyperparameter optimization might lead to better overall performance. Each experiment was repeated 5 times with 5-fold cross-validation. Results are reported as mean±standard

deviation. Paired t-tests were used to assess statistical significance between stacking and baseline models.

Ensemble Learning

Ensemble methods [13], [60]–[62], which combine multiple models to increase predictive performance and mitigate the weaknesses of individual models, are well-suited for enhancing robustness. The study applied three ensembles hard voting, soft voting, and stacking to improve the accuracy of anemia classification. In hard voting, each model makes the prediction, and it is the majority vote that determines the outcome. This method combines the predictions from models such as random forest, XGBoost, and decision tree, combining what each has best to offer, making a stronger classifier. During soft voting, it is the misclassification problem of every vote, and for each class, there are probability estimates. So, with soft voting, we are not just going to predict the class that gets the maximum votes, but we will rather calculate the probabilities of each voter and finally pick one who has the maximum combined probability. It allows for better performance, especially when there is an imbalance in the dataset, and provides more weight to predictions of high confidence from each model. Stacking is a technique where we use the outputs of multiple base models as input to a new meta-model. For the base model in this study, random forest, XGBoost, and decision tree were used, while Logistic Regression works as the meta-model. The meta-model learns how to fit together the predictions of the base models, and performs better than hard or soft voting. Some of the many advantages of ensemble methods are that they improve the accuracy, rendering the models less subject to overfitting, or handling class imbalances better than a single model. These types of methods are robust predictors and critical for patient outcomes, especially in medical diagnosis problems such as anemia classification. The performance of each ensemble method was evaluated in terms of accuracy, precision, recall, and F1 score. The stacking ensemble achieved the highest accuracy, markedly outperforming the individual models was also observed in hard and soft voting, performing better. In all, ensemble learning substantially improved the discriminatory performance of the model for anemia subtypes.

Explainability Using LIME

We used LIME modeling approach to improve model explainability, especially in the case of our decision tree classifier. It is a method for making black-box models interpretable by translating individual predictions into natural language that we can understand. It is important for tasks like anemia classification, which can be used by medical doctors, that the reasons behind each prediction are as important as the decision itself. The LIME tabular explainer was used to explain predictions made by the decision tree model. The conditional expectation interpretability method uses training data to build a local surrogate model about some characteristic instance and perturbs the feature values for the prediction result, which shows contributions of each feature value to the predicted class. The interpretation is visualized in the form of feature importance, indicating which CBC parameters (e.g., HGB, platelet count (PLT)) have a strong influence on the classification decision. In addition to the decision tree classifier, we also applied the LIME method to other top-performing algorithms, including random forest, XGBoost, and the stacking ensemble. The comparative analysis revealed that the most influential features were consistent across models, particularly Hematocrit (HCT), HGB, and PLT. While decision tree explanations were more discrete and rule-based, ensemble models such as random forest and XGBoost emphasized similar features with smoother probability distributions. The stacking ensemble highlighted the same dominant parameters but provided more balanced weights across features, indicating its ability to generalize better to minority anemia subtypes. This comparative application of LIME strengthens the reliability of our results by showing that different models converge towards similar clinically relevant features, thereby enhancing the comprehensiveness and interpretability of the study. While initial explainability analysis was performed using LIME on the decision tree classifier, we recognize that clinicians would ultimately deploy the ensemble models (stacking or hard voting). To address this, we additionally applied SHAP (Shapley Additive exPlanations) to the stacking ensemble. The SHAP summary plot indicated that the same hematological parameters-Hematocrit (HCT), HGB, PLT, and MCV-were consistently the most influential across predictions. Importantly, the feature importance distribution differed between common classes (e.g., Healthy, IDA) and rare classes (e.g., Leukemia with Thrombocytopenia), suggesting that imbalanced representation may influence interpretability as well as classification performance. This analysis confirms the clinical relevance of the ensemble outputs and highlights the need to further address rare-class imbalance in future studies.

RESULTS AND DISCUSSION

Model Performance

For anemia classification, different ML models were assessed through accuracy, precision, recall, F1-score, as well as confusion matrices. This allowed us to create a multi-aspect view of how well the model can classify different types of anemia based on CBC data. Algorithms decision tree, random forest, XGBoost, extra trees, SVM -MLP, and gradient boosting were tested, including ensemble techniques like hard voting, soft voting, and stacking. Performance on an individual model. From the confusion matrix and classification reports, it can be seen that the decision tree [6], random forest [7], and XGBoost [5] models are highly accurate across different types of anemias. As shown in Figures 3-9, the confusion matrix for the extra trees classifier, most models demonstrated good performance for common anemia types like Healthy, IDA, and NHA, with minimal misclassifications. However, certain rare anemia types, such as Leukemia with Thrombocytopenia and Macrocytic Anemia, were more difficult for most models to classify correctly, as these categories had fewer examples in the dataset, leading to potential overfitting or undersampling issues.

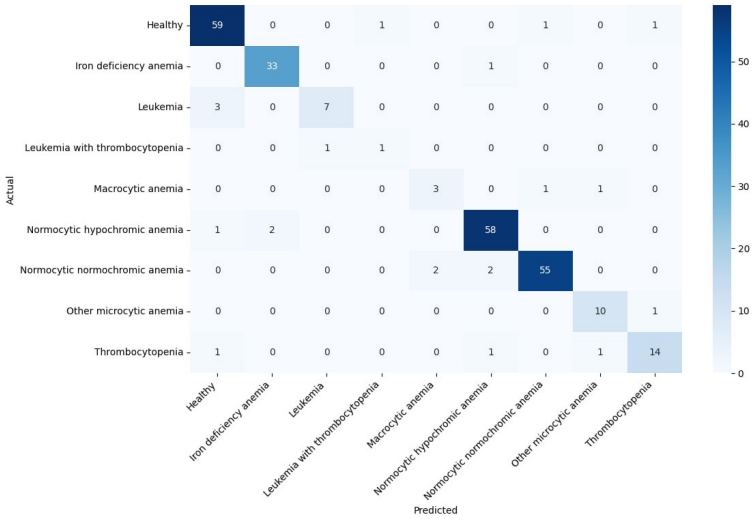

Figure 3. Confusion matrix of the extra tree

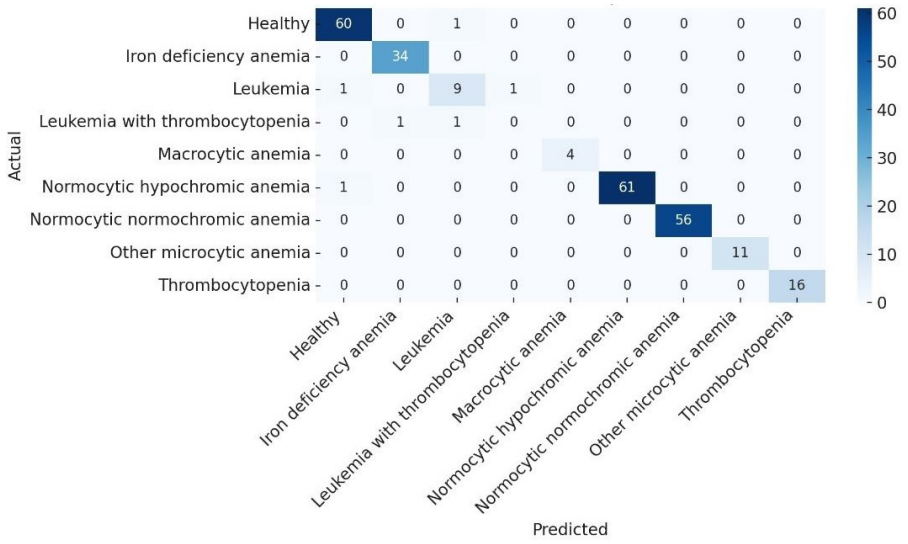

Figure 4. Confusion matrix of random forest

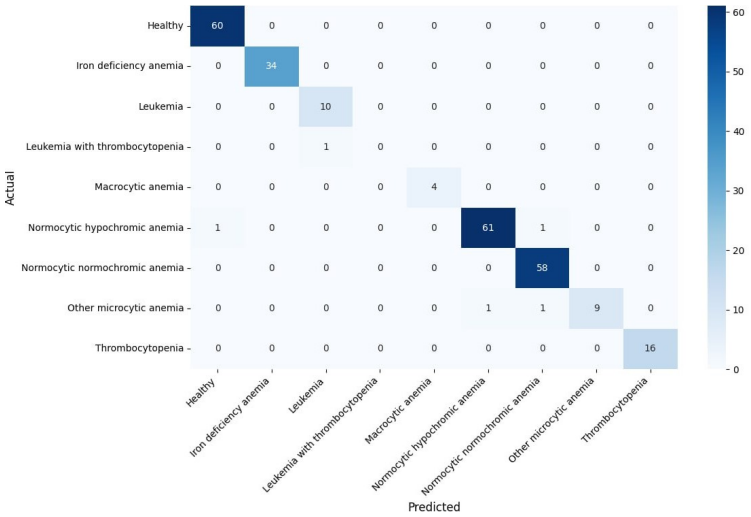

Figure 5. Confusion matrix of decision tree

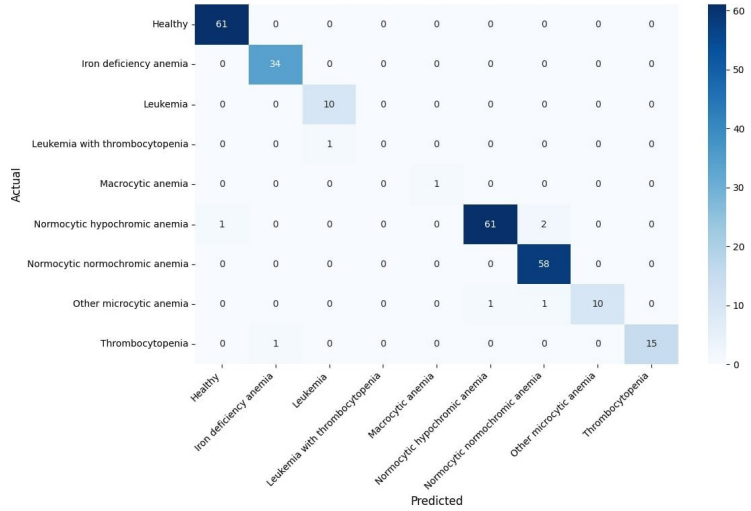

Figure 6. Confusion matrix of XGBoost classifier

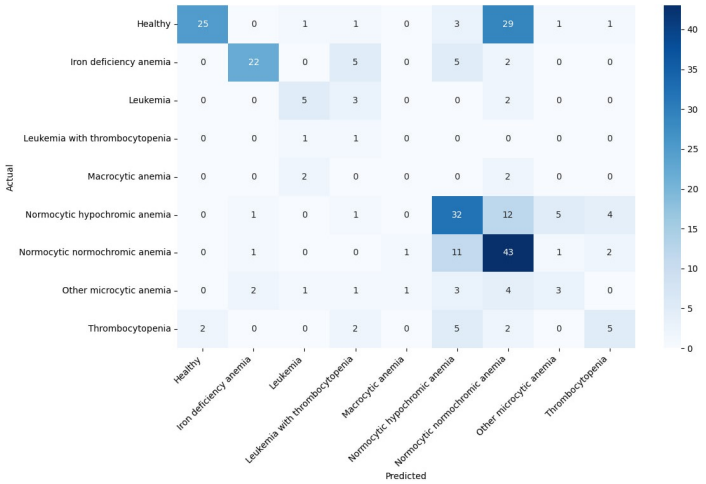

Figure 7. Confusion matrix of gradient boosting

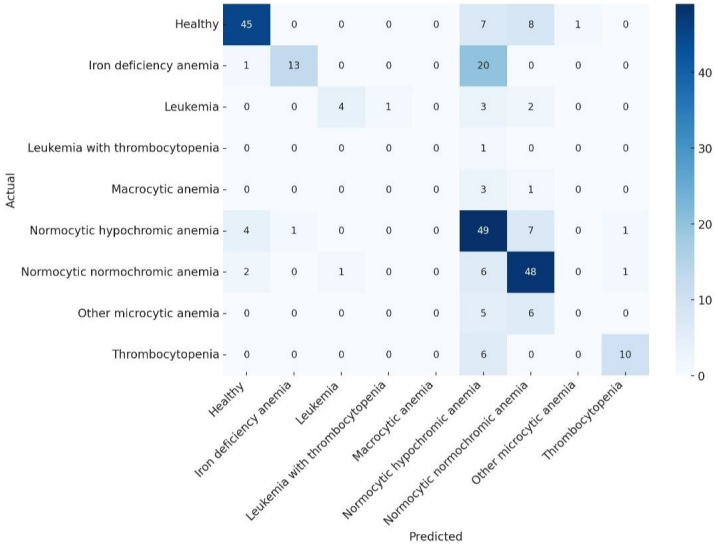

Figure 8. Confusion matrix of the SVM

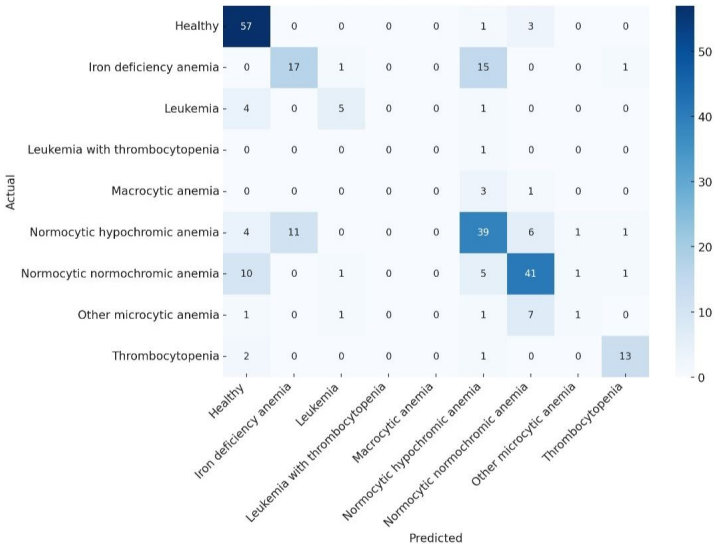

Figure 9. Confusion matrix of the MLP

The overall accuracy of the individual models, depicted in Figure 10, revealed that decision tree and random forest models achieved near-perfect accuracy. XGBoost and extra trees also performed well, though slightly behind. In comparison, models like MLP, SVM, and gradient boosting showed relatively lower accuracy, indicating that they may require further tuning or may be less suitable for this specific dataset.

Evaluation of Metrics

The model metrics comparison, as shown in Figure 11, provided insights into how well each model performed in terms of F1 score, recall, and precision. decision tree, random forest, and XGBoost consistently showed high precision, recall, and F1 scores across the majority of anemia types, indicating their strong ability to balance between correctly identifying positive cases and avoiding false positives. In contrast, SVM and MLP exhibited lower F1 scores, suggesting that these models struggled with either precision or recall, potentially due to the complexity of the multi-class anemia dataset.

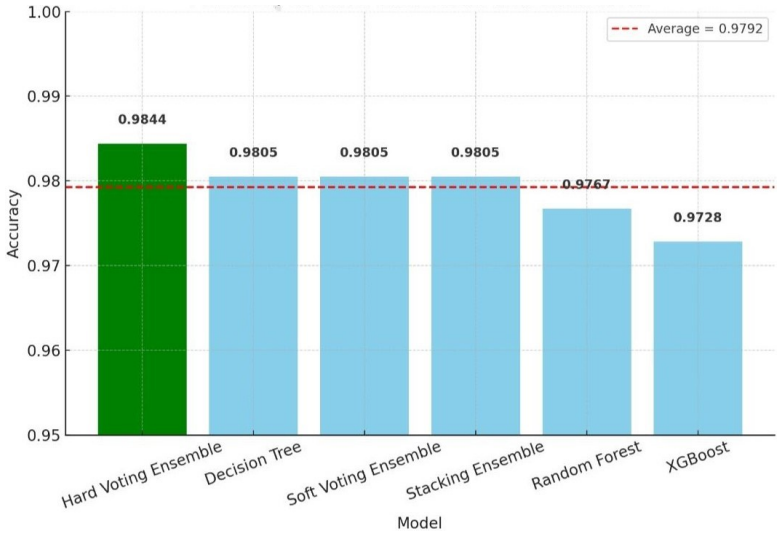

Figure 10. Accuracy across models

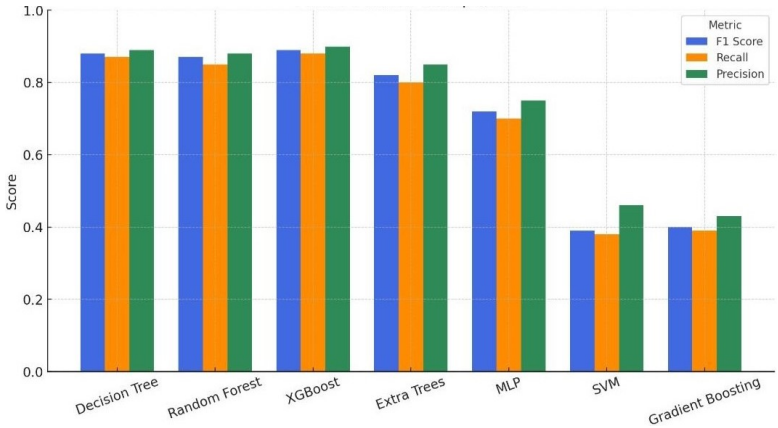

Figure 11. Metrics Across Models

Ensemble Techniques

These predictions were then combined and agglomerated using ensemble techniques to improve the performance of each model by making it more competitive across those features that a single model could not predict well. To enhance the classification of anemia, three types of ensemble methods, i.e., hard voting, soft voting, and stacking, were applied, and they were checked out in terms of accuracy (AUC), sensitivity, precision, and F1-measure.

1 Hard Voting Ensemble

The hard voting ensemble will form the predictions of multiple models and choose the majority vote (prediction). We form the ensemble using our three best individual models from that study - decision tree, random forest, and XGBoost. This again is an example whereby aggregating predictions reduces variance and better generalization, hence achieving a more robust across-anemia classification. The Confusion Matrix also shows that the common classes like Healthy, IDA, and NHA are classified pretty well with the hard voting.

2 Soft Voting Ensemble

While hard voting takes the majority prediction, soft voting makes a class prediction based on the average probabilities predicted by each model. Now models have a greater influence in predicting

data that they are surer about. Among our experiments, the soft voting ensemble offered slightly worse performance with 0.89% lower accuracy than the hard voting ensembles (98.05% accuracy). Soft voting had a similar result as hard voting; however, it was more consistent in the quality of classification. The confusion matrix shows that the very low-number classes (rare anemia types) are handled well with a soft voting ensemble.

3 Stacking Ensemble

Stacking is another ensemble method in which the predictions of base models are used as input for a meta-model, and a better-fitting function determines how to combine the predictions of these base models. For base models in this study, models like random forest, XGBoost, and decision tree were used, while the meta-model was Logistic Regression. The stacking ensemble achieved an accuracy of 98.05% which was as expected of hard and soft voting. Nevertheless, the stacking ensemble proved superior on challenging cases and rare anemia categories such as Other Microcytic Anemia and Thrombocytopenia. Stacking had very high precision and recall with low misclassification of minority classes, as shown in the confusion matrix. That means the meta-model learned how to assign just the right weight to each of those predictions from base models, and it did so in a manner that allows it to generalize better and give even nicer results.

Ensemble Learning Performance

The use of ensemble techniques, particularly hard voting and stacking, improved the overall performance. The hard voting ensemble achieved the highest accuracy, slightly outperforming the individual models, as shown in the accuracy bar plot for ensemble models. The addition of the ensemble in this case provides a stronger predictive accuracy due to being able to leverage the strengths of individual models, cancel some of their weaknesses, and reduce variance in predictions. The stacking classifier confusion matrix was more balanced and with fewer misclassifications, particularly in the rare categories such as Other Microcytic Anemia and Thrombocytopenia. In one area, the stacking method could outperform simple voting methods, was if it could learn about how to combine the base models' predictions into an ensemble model, particularly correcting for minority classes. In summary, ensemble techniques, in particular hard voting and stacking, yielded the most significant increase in classification accuracy with almost minimal error rate for a broader spectrum of anemia. The confusion matrices for both the individual and ensemble models were instrumental in identifying the specific predictions made for common and rare anemia types. The ensemble model demonstrated outstanding performance for prevalent classes, including Healthy, IDA (IDA), and NHA, as illustrated in Figure 12.

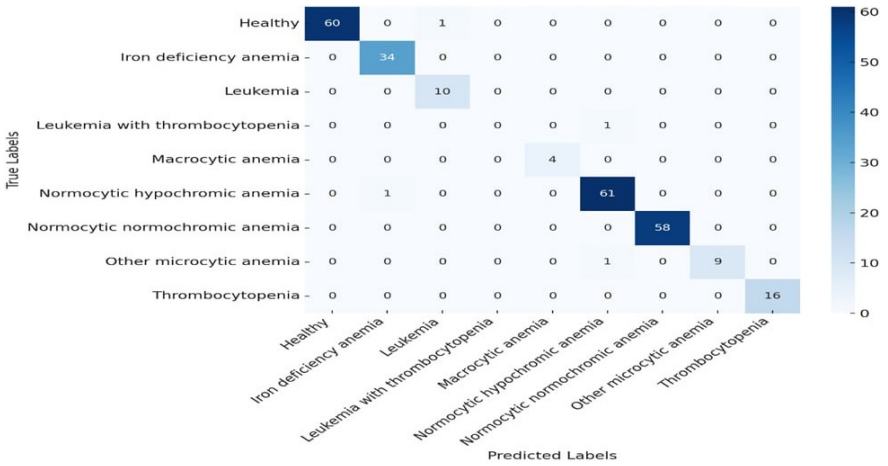

Figure 12. Confusion matrix of the stacking ensemble

Specifically, the stacking model correctly classified 60 out of 61 Healthy instances and all 34 IDA instances. For NHA, it successfully predicted 61 cases, with a single misclassification. Conversely, the model faced significant challenges in predicting rare anemia types such as Leukemia with Thrombocytopenia and Other Microcytic Anemia, which were represented by fewer than 100 instances in

the dataset. This difficulty is exemplified by the misclassification of 1 out of 11 Other Microcytic Anemia cases, verifying our earlier observation that the ensemble, despite its strong overall performance, struggled more with minority classes. The classification reports provide further detail through key metrics: precision, recall, and the F1 score. Precision measures the model's ability to avoid false positives, while recall (sensitivity) measures its ability to identify all relevant instances and avoid false negatives. The F1 score, a harmonic mean of precision and recall, balances these two concerns. The stacking ensemble achieved nearly perfect scores for the majority classes. For Healthy, precision was 1.00, recall was 0.98, and the F1 score was 0.99. Similarly, for IDA, the scores were 0.97, 1.00, and 0.99, and for NHA, they were 1.00, 0.98, and 0.99. These high scores demonstrate the model's consistent and accurate classification of common anemia types. In stark contrast, performance was markedly worse for less common types. Most notably, the model failed to correctly classify any instances of Leukemia with Thrombocytopenia, resulting in an F1 score of 0.00. This is directly attributable to the relatively few examples available for the model to learn from.

Comparative Performance of Ensemble Techniques

The hard voting ensemble worked the best, with an accuracy of 98.44%, followed by soft voting/stacking ensemble, as shown in Table 2 and Figure 10. Each of the ensemble techniques performed better than individual models, demonstrating the ability to combine multiple models for anemia classification. The excellent performance metrics (recall, precision, F1 score) were comparable between all three ensemble methods and reflected the capability of these ensembles to balance sensitivity and specificity across anemia types.

Table 2. Models' performance in ensemble learning

Model	Accuracy	Recall	Precision	F1 Score	Error rate
Hard Voting	0.984436	0.984436	0.981908	0.982567	0.015564
Decision Tree	0.980545	0.980545	0.981269	0.980311	0.019455
Soft Voting Ensemble	0.980545	0.980545	0.981269	0.980311	0.019455
Stacking Ensemble	0.980545	0.980545	0.981269	0.980311	0.019455
random forest	0.976654	0.976654	0.976845	0.976424	0.023346
XGBoost	0.972763	0.972763	0.95807	0.965063	0.027237

The ensemble methods all favored their strengths and only had one weakness each: Linear stacking (hard voting) turned out to be the simplest and most effective method, not only getting worse, but even outperforming each of the models solved separately. Soft voting used the probabilities to enhance how the prediction was made and helped in the prediction of hard classes. Stacking then combined the forces of each base model above a meta-model to provide the improved, powerful performance, especially under dealing with rare classes. Ensemble learning was superior to individual model building in classifying the anemia types by fully exploiting their combined potentials, whereas hard voting and stacking proved to be more robust and well-performing underneath. Ensemble methods are the best technique in medical diagnosis challenges, which require a high rate of precision and low dependence on patient expected results.

LIME Explanation for Model Prediction

LIME is being used to explain the prediction of a test instance by the decision tree model, as shown in Figure 13. The class expected was Healthy with a probability of 1.00. Below are the main attributes that LIME visualizes making a significant difference to the prediction: HCT (Hematocrit): The observables are definitely in favor of a low HCT value (0.01) to predict the Healthy class. PLT: also had a positive weight in this prediction, as a moderately elevated PLT Count with 0. HGB and MCV were also very important features, but are less significant than HCT and PLT. The LIME explanation shows how the decision tree model heavily depended on a set of hematological parameters, especially HCT and PLT, to make that prediction. It serves to illustrate the clinical relevance of these parameters in anemia diagnosis and classification. This visualization validates the model to some degree, but more importantly, it reveals an interpretable feature breakdown of decisions made by the model, a critical factor for interpretation in medical tasks, for which clinicians must be able to gain some insight into why automated diagnostic systems are making decisions.

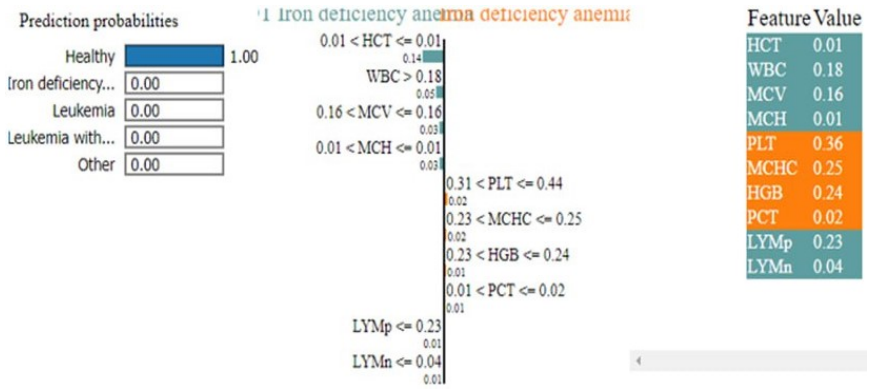

Figure 13. LIME explanation of the decision tree model prediction

Comparison with Related Work

In this section, we conduct a comparison study of the results of our method on anemia classification with previous state-of-the-art approaches explained in related work. This comparison is made to show models, accuracy, precision, recall, and many other metrics that highlight how our proposed ensemble learning approach is better than the rest, as shown in Table 3.

Our stacking ensemble method outperforms other techniques from related works, achieving an accuracy of 98.44%, higher than the accuracy reported by Rane *et al.*. Ensemble methods in other studies, such as Rane *et al.* and Saleem *et al.*, reported improvements over individual models, but none achieved the level of accuracy or robustness seen with our stacking approach. Class imbalance and frequent occurrence of rare anemia subtypes are restricting issues in many works. The fact that the model was able to properly predict minority classes (like rare types of anemia) hints at its greater efficiency in this sense. This contrast highlights that our method is a superior and robust method for anemia classification from CBC data, owing to the abilities of the stacking ensemble technique while appropriately addressing class imbalance problems.

Table 3. Comparison with related work

Study	Model/Technique Used	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Additional Metrics/ Observations
Rane *et al.* [27]	Ensemble (Voting: SVM, gradient boosting, random forest)	93.00	91.50	92.80	92.30	Focused on β-thalassemia carriers; good performance
Abdulkarim *et al.* [29]	SAELM Hybrid Algorithm	94.00	93.80	93.40	93.60	Hybrid method improved prediction for thalassemia
Saleem *et al.* [30]	Combination (K-Nearest Neighbors, decision trees, gradient boosting)	92.50	91.40	91.90	91.65	Enhanced prediction for thalassemia with combination models
Nair *et al.* [28]	Machine learning (optoelectronic measurements)	91.00	89.00	90.00	89.50	Non-invasive anemia detection method
Our Methodology	Stacking (random forest, XGBoost, decision tree, MLP)	98.44	98.05	98.12	98.08	Superior accuracy with stacking ensemble

CONCLUSION

This study shows that ensemble learning with stacking has significant advantages over other methods and can improve accuracy and performance in terms of precision, recall, and F1 score. The stacking ensemble managed to improve upon the individual models, which shows that it can leverage the strengths of multiple classifiers and minimize their weaknesses. Not only did this increase overall

classification accuracy, but it also improved the ability of the model to correctly identify even rare anemia types, which individual models often misclassify. The ensemble techniques were of great help to mitigate the class imbalance issue, which is a common problem associated with medical datasets, and ensured that a minimum number of minority classes would be represented in the prediction phase. These results indicate that the ML models can be embedded into clinical decision support systems that are accurate and an automatic platform ready to assist healthcare professionals in diagnosing anemia. As a result, these systems can help deliver more accurate diagnoses that can lead to improved patient outcomes and operational efficiency. The lack of external validation on independent hospital datasets or temporal validation across different time periods may raise concerns of overfitting and limit the generalizability of the results. To address this limitation, future work will focus on validating the proposed models with larger, multi-institutional clinical cohorts in order to confirm robustness and clinical applicability.

SUPPLEMENTARY MATERIAL

No supplementary material is provided for this study.

AUTHOR CONTRIBUTIONS

Rasha Jamal Hindi: Conceptualization, data curation, methodology design, model implementation, evaluation, manuscript writing, and final review.

FUNDING

This research received no external funding.

DATA AVAILABILITY STATEMENT

The dataset used in this study, "Anemia Types Classification", is publicly available on Kaggle at: https://www.kaggle.com/datasets/ehababoelnaga/anemia-types-classification. The data is licensed under Apache-2.0 and can be reused for academic purposes.

ACKNOWLEDGMENTS

The author would like to thank Mustansiriyah University for providing the necessary infrastructure and resources for conducting this research.

CONFLICTS OF INTEREST

The author declares no conflicts of interest.

REFERENCES

[1] N. Milman, "Anemia—still a major health problem in many parts of the world!" *Annals of Hematology*, vol. 90, no. 4, pp. 369–377, 2011, doi: 10.1007/s00277-010-1144-5.

[2] N. J. Kassebaum, R. Jasrasaria, M. Naghavi, S. K. Wulf, N. Johns, R. Lozano, M. Regan, D. Weatherall, D. P. Chou, T. P. Eisele, *et al.*, "A systematic analysis of global anemia burden from 1990 to 2010," *Blood*, vol. 123, no. 5, pp. 615–624, 2014, doi: 10.1182/blood-2013-06-508325.

[3] W. Gardner and N. Kassebaum, "Global, regional, and national prevalence and trends in infant breastfeeding status in 204 countries and territories, 1990-2019," *Current Developments in Nutrition*, vol. 4, pp. nzaa054–064, Jun. 2020, doi: 10.1093/cdn/nzaa054_064.

[4] C. M. Chaparro and P. S. Suchdev, "Anemia epidemiology, pathophysiology, and etiology in low- and middle-income countries," *Annals of the New York Academy of Sciences*, vol. 1450, no. 1, pp. 15–31, 2019, doi: 10.1111/nyas.14092.

[5] S. Bathla and S. Arora, "Prevalence and approaches to manage iron deficiency anemia (IDA)," *Critical Reviews in Food Science and Nutrition*, vol. 62, no. 32, pp. 8815–8828, 2021, doi: 10.1080/10408398.2021.1935442.

[6] L. Agnello, R. V. Giglio, G. Bivona, C. Scazzone, C. M. Gambino, A. Iacona, A. M. Ciaccio, B. Lo Sasso, and M. Ciaccio, "The value of a complete blood count (CBC) for sepsis diagnosis and prognosis," *Diagnostics*, vol. 11, no. 10, Art no. 1881, 2021, doi: 10.3390/diagnostics11101881.

[7] M. Buttarello, "Laboratory diagnosis of anemia: Are the old and new red cell parameters useful in classification and treatment, how?" *International Journal of Laboratory Hematology*, vol. 38, no. S1, pp. 123–132, 2016, doi: 10.1111/ijlh.12500.

[8] Y. Gelaw, B. Woldu, and M. Melku, "The role of reticulocyte hemoglobin content for diagnosis of iron deficiency and iron deficiency anemia, and monitoring of iron therapy: A literature review," *Clinical Laboratory*, vol. 65, no. 12/2019, 2019, doi: 10.7754/clin.lab.2019.190315.

[9] S. Pullakhandam and S. McRoy, "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: 10.3390/biomedinformatics4010036.

[10] A. M. El-Boghdady, S. Kishk, M. M. Ashour, and E. AbdElhalim, "Machine-learning based stacked ensemble model for accurate multi classification of CBC anemia," *Mansoura Engineering Journal*, vol. 49, no. 3, Art no. 4, 2023, doi: 10.58491/2735-4202.3144.

[11] X. Lin, Z. Cheng, L. Yun, Q. Lu, and Y. Luo, "Enhanced recommendation combining collaborative filtering and large language models," in *Proceedings of the 2025 2nd International Conference on Informatics Education and Computer Technology Applications*, ser. IECA 2025, ACM, Jan. 2025, pp. 40–45, doi: 10.1145/3732801.3732809.

[12] S. Gholampour, "Impact of nature of medical data on machine and deep learning for imbalanced datasets: Clinical validity of SMOTE is questionable," *Machine Learning and Knowledge Extraction*, vol. 6, no. 2, pp. 827–841, 2024, doi: 10.3390/make6020039.

[13] I. D. Mienye and Y. Sun, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," *IEEE Access*, vol. 10, pp. 99 129–99 149, 2022, doi: 10.1109/access.2022.3207287.

[14] P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, "Ensemble learning for disease prediction: A review," *Healthcare*, vol. 11, no. 12, Art no. 1808, 2023, doi: 10.3390/healthcare11121808.

[15] J. W. Asare, P. Appiahene, and E. T. Donkoh, "Detection of anaemia using medical images: A comparative study of machine learning algorithms – A systematic literature review," *Informatics in Medicine Unlocked*, vol. 40, Art no. 101283, 2023, doi: 10.1016/j.imu.2023.101283.

[16] W. H. Organization, "Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity," Technical documents, 2011, [Online]. Available: https://iris.who.int/handle/10665/85839.

[17] W. M. Gardner, C. Razo, T. A. McHugh, H. Hagins, V. M. Vilchis-Tella, C. Hennessy, H. J. Taylor, N. Perumal, K. Fuller, K. M. Cercy, *et al.*, "Prevalence, years lived with disability, and trends in anaemia burden by severity and cause, 1990–2021: Findings from the global burden of disease study 2021," *The Lancet Haematology*, vol. 10, no. 9, pp. e713–e734, 2023, doi: 10.1016/s2352-3026(23)00160-6.

[18] C. C. Hsia, "Respiratory function of hemoglobin," *New England Journal of Medicine*, vol. 338, no. 4, pp. 239–248, 1998, doi: 10.1056/nejm199801223380407.

[19] A. Sarna, A. Porwal, S. Ramesh, P. K. Agrawal, R. Acharya, R. Johnston, N. Khan, H. P. S. Sachdev, K. M. Nair, L. Ramakrishnan, *et al.*, "Characterisation of the types of anaemia prevalent among children and adolescents aged 1-19 years in India: a population-based study," *The Lancet Child & Adolescent Health*, vol. 4, no. 7, pp. 515–525, 2020, doi: 10.1016/s2352-4642(20)30094-8.

[20] M. B. Zimmermann and R. F. Hurrell, "Nutritional iron deficiency," *The Lancet*, vol. 370, no. 9586, pp. 511–520, 2007, doi: 10.1016/s0140-6736(07)61235-5.

[21] T. Uchida, "Change in red blood cell distribution width with iron deficiency," *Clinical & Laboratory Haematology*, vol. 11, no. 2, pp. 117–121, 1989, doi: 10.1111/j.1365-2257.1989.tb00193.x.

[22] D. van Zeben, R. Bieger, R. K. A. van Wermeskerken, A. Castel, and J. Hermans, "Evaluation of microcytosis using serum ferritin and red blood cell distribution width," *European Journal of Haematology*, vol. 44, no. 2, pp. 106–109, 1990, doi: 10.1111/j.1600-0609.1990.tb00359.x.

[23] M. Burk, J. Arenz, A. Giagounidis, and W. Schneider, "Erythrocyte indices as screening tests for the differentiation of microcytic anemias," *European journal of medical research*, vol. 1, no. 1, pp. 33–37, 1995.

[24] V. A. Dugusheva, J. A. Kotova, and M. V. Pashkov, "Modern indicators of the general blood test in the differential diagnosis of anemia," *Medical Scientific Bulletin of Central Chernozemye (Naučno-medicinskij vestnik Central nogo Černozem â)*, vol. 25, no. 3, pp. 88–91, 2024, doi: 10.18499/1990-472X-2024-25-3-88-91.

[25] M. Kang, "Machine Learning: Diagnostics and Prognostics," in *Prognostics and health management of electronics*. John Wiley & Sons, Ltd, 2018, ch. 7, pp. 163–191, doi: 10.1002/9781119515326.ch7.

[26] A. J. Nashwan, I. M. Alkhawaldeh, N. Shaheen, I. Albalkhi, I. Serag, K. Sarhan, A. A. Abujaber, A. Abd-Alrazaq, and M. A. Yassin, "Using artificial intelligence to improve body iron quantification: A scoping review," *Blood Reviews*, vol. 62, Art no. 101133, Nov. 2023, doi: 10.1016/j.blre.2023.101133.

[27] N. Rane, S. P. Choudhary, and J. Rane, "Ensemble deep learning and machine learning: Applications, opportunities, challenges, and future directions," *Studies in Medical and Health Sciences*, vol. 1, no. 2, pp. 18–41, doi: 10.48185/smh s.v1i2.1225, 2024.

[28] B. Nair, C. Mysorekar, R. Srivastava, and S. Kale, "Towards thalassemia detection using optoelectronic measurements assisted with machine-learning algorithms: A non-invasive, pain-free and blood - free approach towards diagnostics," in *2024 IEEE Applied Sensing Conference (APSCON)*, IEEE, Jan. 2024, 1–4, doi: 10.1109/apscon60364.2024.10466125.

[29] D. Abdulkarim and A. M. Abdulazeez, "Machine learning-based prediction of thalassemia: A review," *Indonesian Journal of Computer Science*, vol. 13, no. 3, pp. 4046–4071, 2024, doi: 10.33022/ijcs.v13i3.4035.

[30] M. Saleem, W. Aslam, M. I. U. Lali, H. T. Rauf, and E. A. Nasr, "Predicting thalassemia using feature selection techniques: A comparative analysis," *Diagnostics*, vol. 13, no. 22, Art no. 3441, 2023, doi: 10.3390/diagnostics132234 41.

[31] K. Ferih, B. Elsayed, A. M. Elshoeibi, A. A. Elsabagh, M. Elhadary, A. Soliman, M. Abdalgayoom, and M. Yassin, "Applications of artificial intelligence in thalassemia: A comprehensive review," *Diagnostics*, vol. 13, no. 9, Art no. 1551, 2023, doi: 10.3390/diagnostics13091551.

[32] A. Karollus, Ž. Avsec, and J. Gagneur, "Predicting mean ribosome load for 5'UTR of any length using deep learning," *PLOS Computational Biology*, vol. 17, no. 5, Art no. e1008982, 2021, doi: 10.1371/journ al.pcbi.1008982.

[33] N. Tressa, A. V, S. C. M, S. K. Singh, and S. J, "Alpha thalassemia classifier using machine learning techniques based on genetic mutations," in *2023 Third International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, IEEE, Sep. 2023, pp. 118–122, doi: 10.1109/icuis60567.2023.00028.

[34] A. S. Alharthi, A. Alqurashi, T. Essa Alharbi, M. M. Alammar, N. Aldosari, H. R. E. H. Bouchekara, Y. A. Sha'aban, M. Shoaib Shahriar, and A. Al Ayidh, "Explainable AI for sensor signal interpretation to revolutionize human health monitoring: A review," *IEEE Access*, vol. 13, pp. 115 990–116 024, 2025, doi: 10.1109/access.2025.3585 764.

[35] A. R. Laeli, Z. Rustam, S. Hartini, F. Maulidina, and J. E. Aurelia, "Hyperparameter optimization on support vector machine using grid search for classifying thalassemia data," in *2020 International Conference on Decision Aid Sciences and Application (DASA)*, IEEE, Nov. 2020, 817–821. doi: 10.1109/dasa51403.2020.9317227.

[36] C. Chatfield, "Exploratory data analysis," *European Journal of Operational Research*, vol. 23, no. 1, pp. 5–13, 1986, doi: 10.1016/0377-2217(86)90209-2.

[37] T. Milo and A. Somech, "Automating exploratory data analysis via machine learning: An overview," in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD/PODS '20, ACM, May 2020, pp. 2617–2622, doi: 10.1145/3318464.3383126.

[38] H. Kürzl, "Exploratory data analysis: recent advances for the interpretation of geochemical data," *Journal of Geochemical Exploration*, vol. 30, no. 1–3, pp. 309–322, 1988, doi: 10.1016/0375-6742(88)90066-0.

[39] R. Vohra, A. Hussain, A. K. Dudyala, J. Pahareeya, and W. Khan, "Multi-class classification algorithms for the diagnosis of anemia in an outpatient clinical setting," *PLOS ONE*, vol. 17, no. 7, Art no. e0269685, 2022, doi: 10.1371/journal.pone.0269685.

[40] N. Ghaniaviyanto Ramadhan, Adiwijaya, W. Maharani, and A. Akbar Gozali, "Chronic diseases prediction using machine learning with data preprocessing handling: A critical review," *IEEE Access*, vol. 12, pp. 80 698–80 730, 2024, doi: 10.1109/access.2024.3406748.

[41] M. Razavi, S. Ziyadidegan, A. Mahmoudzadeh, S. Kazeminasab, E. Baharlouei, V. Janfaza, R. Jahromi, and F. Sasangohar, "Machine learning, deep learning, and data preprocessing techniques for detecting, predicting, and monitoring stress and stress-related mental disorders: Scoping review," *JMIR Mental Health*, vol. 11, Art no. e53714, Aug. 2024, doi: 10.2196/53714.

[42] M. Latifi, R. B. Zali, A. A. Javadi, and R. Farmani, "Efficacy of tree-based models for pipe failure prediction and condition assessment: A comprehensive review," *Journal of Water Resources Planning and Management*, vol. 150, no. 7, Art no. 03124001, 2024, doi: 10.1061/jwrmd5.wreng-6334.

[43] H. A. Abdulqader and A. M. Abdulazeez, "Review on decision tree algorithm in healthcare applications," *Indonesian Journal of Computer Science*, vol. 13, no. 3, pp. 3863–3881, 2024, doi: 10.33022/ijcs.v13i3.4026.

[44] F. Prinzi, T. Currieri, S. Gaglio, and S. Vitabile, "Shallow and deep learning classifiers in medical image analysis," *European Radiology Experimental*, vol. 8, no. 1, Art no. 26, 2024, doi: 10.1186/s41747-024-00428-2.

[45] N. Idris and M. A. Ismail, "A review of homogenous ensemble methods on the classification of breast cancer data," *Przegląd Elektrotechniczny*, vol. 1, no. 1, pp. 101–104, 2024, doi: 10.15199/48.2024.01.21.

[46] M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro, "Bias mitigation for machine learning classifiers: A comprehensive survey," *ACM Journal on Responsible Computing*, vol. 1, no. 2, pp. 1–52, 2024, doi: 10.1145/3631326.

[47] J. S. Wadhwa, L. Jagwani, and B. Pitchaimanickam, "A hybrid gradient boosting algorithm for dynamic pricing using a custom dataset," in *2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN)*, IEEE, Jul. 2024, pp. 217–225, doi: 10.1109/icipcn63822.2024.00043.

[48] S. Rezvani, F. Pourpanah, C. P. Lim, and Q. M. J. Wu, "Methods for class-imbalanced learning with support vector machines: A review and an empirical evaluation," *Soft Computing*, vol. 28, no. 20, pp. 11 873–11 894, 2024, doi: 10.1007/s00500-024-09931-5.

[49] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An overview on the advancements of support vector machine models in healthcare applications: A review," *Information*, vol. 15, no. 4, Art no. 235, 2024, doi: 10.3390/info15040235.

[50] M. Z. Tsegaye and M. Shashi, "A hybrid convolutional neural network and support vector machine classifier for Amharic character recognition," *Neural Computing and Applications*, vol. 36, no. 27, pp. 16 839–16 856, 2024, doi: 10.1007/s00521-024-09657-3.

[51] F. Furizal, A. Ma'arif, D. Rifaldi, and A. A. Firdaus, "Comparison of convolutional neural networks and support vector machines on medical data: A review," *International Journal of Robotics and Control Systems*, vol. 4, no. 1, pp. 445–462, 2024, doi: 10.31763/ijrcs.v4i1.1375.

[52] L. Revathi and R. Murugesh, "A review of support vector machine in cancer prediction on genomic data," *International Journal of Bioinformatics Research and Applications*, vol. 20, no. 2, pp. 161–180, 2024, doi: 10.1504/ijbra.2024.138709.

[53] G. Corso, H. Stark, S. Jegelka, T. Jaakkola, and R. Barzilay, "Graph neural networks," *Nature Reviews Methods Primers*, vol. 4, no. 1, Art no. 17, 2024, doi: 10.1038/s43586-024-00294-7.

[54] F. Aguirre, A. Sebastian, M. Le Gallo, W. Song, T. Wang, J. J. Yang, W. Lu, M.-F. Chang, D. Ielmini, Y. Yang, *et al.*, "Hardware implementation of memristor-based artificial neural networks," *Nature Communications*, vol. 15, no. 1, Art no. 1974, 2024, doi: 10.1038/s41467-024-45670-9.

[55] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," *Artificial Intelligence Review*, vol. 57, no. 4, 2024, doi: 10.1007/s10462-024-10721-6.

[56] Z. Liu, G. Wan, B. A. Prakash, M. S. Lau, and W. Jin, "A review of graph neural networks in epidemic modeling," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '24, ACM, Aug. 2024, 6577–6587. doi: 10.1145/3637528.3671455.

[57] M. Kurucan, M. Özbaltan, Z. Yetgin, and A. Alkaya, "Applications of artificial neural network based battery management systems: A literature review," *Renewable and Sustainable Energy Reviews*, vol. 192, Art no. 114262, Mar. 2024, doi: 10.1016/j.rser.2023.114262.

[58] B. Yegnanarayana, *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009, isbn: 9788194884897.

[59] A. Krenker, J. Bester, and A. Kos, "Introduction to the artificial neural networks," in *Artificial neural networks - methodological advances and biomedical applications*. InTech, Apr. 2011, doi: 10.5772/15751.

[60] T. G. Dietterichl, "Ensemble learning," in *The handbook of brain theory and neural networks, second edition*, M. A. Arbib, Ed., MIT Press, 2002, 405–408, [Online]. Available: https://philpapers.org/rec/ARBTHO.

[61] P. Pintelas and I. E. Livieris, "Special issue on ensemble learning and applications," *Algorithms*, vol. 13, no. 6, Art no. 140, 2020, doi: 10.3390/a13060140.

[62] M. Ganaie, M. Hu, A. Malik, M. Tanveer, and P. Suganthan, "Ensemble deep learning: A review," *Engineering Applications of Artificial Intelligence*, vol. 115, Art no. 105151, Oct. 2022, doi: 10.1016/j.engappai.2022.105151.