

Sequential Random Split with Random Feature Subsetting for imbalanced Iron Deficiency Anemia Classification: A Machine Learning Study

S. N. Lakshmi Malluvalasa*

Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, A.P., India

E-mail: sweacha.lakshmi@gmail.com

ORCID iD: <https://orcid.org/0000-0002-9399-5048>

*Corresponding Author

Sajana T.

Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, A.P., India

E-mail: sajana.cse@kluniversity.in

ORCID iD: <https://orcid.org/0000-0003-0317-0892>

Received: 11 June, 2026; Revised: 25 June, 2026; Accepted: 20 July, 2026; Published: 08 August, 2026

Abstract: Iron Deficiency Anemia (IDA) is the most common type of anemia and can adversely affect quality of life by reducing oxygen delivery to body tissues. Accurate diagnosis is essential but remains challenging due to the manual interpretation of Complete Blood Count (CBC) parameters, which can be time-consuming and susceptible to human error. Although machine learning models have explored for early IDA detection, imbalanced datasets often lead to biased predictions and reduced performance for minority classes. Furthermore, conventional feature selection approaches may face challenges when dealing with high-dimensional medical data, potentially affecting predictive performance and model robustness. This study proposed a Sequential Random Split (SRS) framework combined with Random Feature Subsetting strategy for iron deficiency anemia (IDA) classification. The framework explores diverse feature combinations and evaluates their impact on predictive performance using a real-world imbalanced IDA dataset. The proposed approach compared with few ensemble learning methods, including Boosting, Stacking, Random Subspace, Random Patches, Extra Trees, Voting, Bagging, and Gradient Boosting. Model performance assessed using Accuracy, Precision, Recall and F1-Score metrics. The proposed SRS framework achieved the highest classification performance among the evaluated ensemble learning approaches. For class-1 (IDA), the model achieved an accuracy of 94.11%, F1-Score of 96%, and Precision of 96%. For Class-0 (non-IDA), the model achieved a Precision of 88%, F1-Score of 88%, and recall of 88%. These findings indicate that the proposed SRS framework provides strong classification performance on the imbalanced IDA dataset. The experimental results indicate that the proposed SRS framework achieved competitive performance for IDA classification on the evaluated dataset. The combination of Sequential random split and Random Feature Subsetting contributed to improved predictive performance compared with the evaluated ensemble learning approaches. These findings suggest that the proposed framework may serve as a useful machine learning strategy for supporting IDA diagnosis, while further validation on larger and more diverse clinical datasets is required to assess its generalizability and robustness.

Index Terms: Iron Deficiency Anemia, Anemia, Classification, Machine Learning, Imbalanced Data, Ensemble learning, Diagnostics

1. Introduction

Anemia, one of the most common blood illnesses, occurs when red blood cells or hemoglobin levels are low and it leads to reduction of oxygen supply to tissues, affects more people. Most blood issues caused by gene abnormalities, which can inherit. Medical disorders like drug use and lifestyle decisions might cause blood abnormalities. According to research, anemia is the most frequent blood condition [1,2]. It produces weakness, fatigue, dizziness, and shortness of breath. Age, sex, smoking, and pregnancy affect the optimum hemoglobin concentration and anemia most seen is Iron Deficiency Anemia (IDA) [3,4,5].

According to the World Health Organization, iron deficiency anaemia characterized by a hemoglobin concentration of 110 gm/l at sea level in children under the age of five and pregnant women, and a haemoglobin concentration of 120 gm/l in non-pregnant women. Anaemia affects over half a billion women 15-49 years of age and 269 million children 6-59 months of age worldwide. In 2019, 30% (539 million) of non-pregnant women and 37% (32 million) of pregnant women aged 15-49 years affected by anemia [6]. The prevalence of anemia is especially high in developing nations [7,8]. Anemia significantly affects quality of life and can occur either as a primary disease or as a symptom associated with more serious medical conditions [9,10]. Therefore, accurate diagnosis and timely management are important in any clinical setting.

Individuals who suffer from iron deficiency anemia, vitamin B12 deficiency anemia, and anemia caused by chronic disease are among the most frequent types of anemia and to diagnose this Complete blood count(CBC) parameters are manually interpreted in diagnostic procedure which leads to time consuming process and prone to human error [11,12,13,14]. In recent years, predictive models have been applied to enhance the early detection of IDA. These models aim to provide rapid, non-invasive, and accurate screening tools by leveraging clinical and non-clinical data. Traditional classifiers such as Support Vector Machines (SVMs), Random Forests, and logistic regressions have commonly used due to their robustness in handling high-dimensional medical data. Deep learning models including convolutional neural networks (CNNs) and Recurrent Neural Networks(RNNs) have also been employed for complex feature extraction and temporal analysis in longitudinal health records. Previous studies primarily utilized CBC parameters, demographic factors (age, gender), and clinical history as predictive features with few approaches integrating non-invasive features such as dietary intake patterns, symptom analysis, and wearable device data for real-time monitoring.

Recent advances have further demonstrated the effectiveness of machine learning for anemia diagnosis and clinical decision support. Comparative studies have shown that ensemble learning and deep learning models can achieve high predictive performance for anemia detection while reducing the need for invasive diagnostic procedures [38,39]. Furthermore, explainable artificial intelligence (XAI) techniques have recently incorporated into machine learning models to improve transparency and clinical interpretability, thereby increasing their potential for real-world healthcare applications [39]. These findings highlight the growing importance of robust feature selection, imbalance handling, and interpretable machine learning models for reliable anemia classification [40].

However, constructing accurate predictive models for iron deficiency anemia remains challenging due to the imbalance in datasets, where non-anemic samples significantly outnumber anemic samples. This imbalance can result in biased models favouring the majority class and reducing the model's effectiveness. To address this, traditional approaches have employed oversampling and undersampling techniques such as Synthetic Minority Over Sampling Techniques (SMOTE) and Random under sampling as well as advanced methods like Cost-sensitive learning and Ensemble techniques. Despite these efforts, most models face challenges in maintaining generalizability across diverse populations and lack interpretability, which limits clinical adoption. Moreover, existing studies rely on conventional feature selection methods like recursive feature Elimination and Principal Component analysis which limited in capturing complex feature interactions in high dimensional medical data.

Although Random Subspace, Bagging, and Random Forest methods utilize random feature selection to increase model diversity, the proposed Sequential Random Split (SRS) framework differs in its implementation and objective. Random Forest performs feature sampling internally during tree construction, while Random Subspace generates feature subsets primarily to improve ensemble diversity. In contrast, the proposed SRS framework evaluates multiple feature subsets across sequential data partitions to investigate their influence on classification performance in imbalanced iron deficiency anemia data. Therefore, the proposed framework should regard as a feature exploration and evaluation strategy rather than a modification of tree-based ensemble learning algorithms.

To overcome these limitations, this study proposes a Sequential Random Split (SRS) framework combined with Random Feature Subsetting strategy for Iron Deficiency Anemia (IDA) classification on imbalanced clinical data. The Random Feature Subsetting component within the SRS framework explores diverse feature combinations that may assist machine learning models in identifying relationships among these hematological parameters relevant to IDA classification. By applying Random Feature Subsetting within the Sequential Random Split (SRS) framework to an imbalanced IDA dataset, this study aims to improve predictive performance and model robustness. Additionally, the proposed framework compared with ensemble learning approaches to assess its impact on classification performance and generalizability. The findings may support the development of machine learning-based decision support tools for IDA screening and contribute to medical informatics.

This study focuses on the binary classification of iron deficiency anemia using routinely collected Complete Blood Count (CBC) parameters. The proposed framework assumes that the available hematological indicators contain sufficient discriminatory information for machine learning-based classification. The objective of this work is to investigate whether Sequential Random Split (SRS) combined with Random Feature Subsetting can improve predictive performance and model robustness on an imbalanced clinical dataset. The study does not introduce a new optimization algorithm, loss function, or theoretical learning framework; rather, it evaluates a practical machine learning strategy for feature exploration and classification in the context of IDA detection.

2. Related Work

In medical diagnosis, imbalanced datasets present a challenge, particularly for disorders with low prevalence, which hampers machine learning performance. Techniques like oversampling, undersampling, and synthetic data generation methods such as SMOTE [15] have widely explored to address this issue by balancing class distribution and improving predictive model performance. Feature selection also plays a key role in enhancing model accuracy and interpretability by identifying the most relevant features. Traditional methods like Recursive Feature Elimination (RFE) and Lasso commonly used to reduce dimensionality [16]. More recently, Random Subset Selection (RSS) has gained attention for handling high-dimensional data by building models from randomly selected feature subsets to identify the most informative variables [17]. In iron deficiency research, studies have focused on identifying key biomarkers and building predictive models. For example, combined clinical and biochemical parameters to predict iron deficiency anemia in children, while applied machine learning techniques to predict iron deficiency in pregnant women, underscoring the importance of early detection and intervention [18,19,20]. Moreover, recent advancements in ensemble learning methods such as Random Forests [21] and Gradient Boosting Machines [22] have shown promise in improving the performance of models trained on imbalanced datasets. For example, [23] demonstrated the effectiveness of combining SMOTE with Random Forests for predicting rare medical conditions, indicating the potential benefits of such hybrid approaches.

Recent studies have continued to improve anemia classification by integrating ensemble learning, feature selection, and explainable artificial intelligence techniques. Appiahene et al. demonstrated that conventional machine learning and deep learning algorithms can accurately detect anemia while reducing reliance on invasive diagnostic procedures [38]. Asare et al. further compared multiple machine learning models using fingernail, palm, and conjunctiva images and reported that ensemble-based approaches achieved reliable diagnostic performance for iron deficiency anemia detection [40]. These recent studies emphasize that effective data preprocessing, feature selection, and ensemble learning remain fundamental components for developing accurate and clinically interpretable diagnostic models [39].

Anomaly detection techniques offer a promising approach to handling imbalanced datasets, as minority class instances can be treated as anomalies. Methods such as one-class SVM [24] and isolation forests [25] have been used to enhance the detection of rare events. In healthcare, machine learning algorithms are widely employed to predict diseases [26], with successful models aiding in diagnosing iron deficiency anemia [27]. In [28], a system using complete blood count reports and the C4.5 decision tree, along with support vector machines (SVM) via the WEKA tool, classified anemia severity. The C4.5 decision tree achieved a 99.42% success rate, outperforming the SVM's 88.13% success rate.

The study in [29] utilized full blood count parameters such as WBC, RBC, HGB, HCT, MCV, MCHC, PLT, NEUT, LYMP, MONO, EO, BO, along with age, as input data. Classifiers like J48 decision tree, multi-layered perceptron, and Naïve Bayes were employed, achieving success rates of 97.16%, 86.55%, and 70.28%, respectively. Machine learning tools such as SVMs, Naïve Bayes, decision trees, k-nearest neighbours, multilayer perceptron's, hybrid classifiers, and genetic algorithm-based models have been widely used to classify different forms of anemia [30,31,32,33,34,35], with hemoglobin level estimation being a key step [36]. Given the advances in machine learning, this study will apply a new model, Sequential Random Split (SRS), and compare it with various ensemble models to determine the most accurate approach for forecasting Iron Deficiency Anemia (IDA) efficiently.

Recent studies have increasingly applied machine learning and ensemble learning techniques to healthcare datasets characterized by class imbalance. [26] demonstrated the effectiveness of multi-class machine learning models for anemia diagnosis in clinical settings. [30] employed machine learning approaches for anemia diagnosis and prediction, highlighting the potential of data-driven clinical decision support systems. Similarly, [34] reported encouraging results using machine learning algorithms for anemia prediction. These studies indicate that robust feature selection, class imbalance handling, and ensemble learning strategies remain important research directions for improving predictive performance in hematological disease classification.

3. Materials and Methods

3.1. Data Set and its description

This study utilizes a publicly available dataset obtained from Mendeley Data [37], specifically the dataset titled "Dataset for Anemia Classification Based on Hematological Parameters" (DOI: 10.17632/dy9mfjchm7.1). The dataset comprises hematological parameters essential for diagnosing and classifying anemia, making it well-suited for this research on identifying iron deficiency anemia (IDA) using machine learning techniques, and complete sample count represented in the below table 1. This rich dataset is the core of our study, giving the required information to create and evaluate machine learning models for anemia detection.

Table 1. Data Set Parameters and description

Parameter	Description	Unit	Reference Range
AGE	AGE	Years	Varies with age group
SEX	Male/Female	1-0	
RBC	Red Blood Cells	Million/	4.44-5.61
HGB	Hemoglobin	g/dL	13.5-16.9
PCV	Packed cell volume	%	40-36
MCV	Mean Corpuscular Volume	fL	81.8-95.5
MCHC	Mean Corpuscular Hemoglobin Concentration	pg	27-32.3
MCH	Mean Corpuscular Hemoglobin	g/dL	32.35
RDW	Red Cell Distribution Width	%	12-13.6
TLC	Total Leukocyte Count	$\times 10^3$ cells/ μ L	4.0
PLT	Platelet Count	$\times 10^3/\text{mm}^3$	150

3.2. Methodology: Sequential Random Split

Accurate detection of iron deficiency anemia using clinical CBC data presents two key challenges:

- Data Partitioning robustness: Different train-test partitions may lead to variability in model performance estimation.
- Feature Overfitting: Static feature selection may expose the model to repetitive feature combinations., increasing the risk of overfitting and reducing generalization capability.

To address these challenges, we propose a Sequential Random Split framework with Random Feature Subsetting strategy. This dual randomization strategy integrates:

- Sequential Random Split: Generates randomized training and testing partitions while maintaining consistent data partitioning across repeated experiments. The framework introduces controlled randomness to improve the robustness and reproducibility of model evaluation.
- Random Feature Subset Selection: Dynamically varies the input feature subsets during each iteration, enhancing model generalization and reducing the risk of overfitting.

This approach is particularly suited for clinical data analysis involving imbalanced datasets, where robust data partitioning and diverse feature combinations are important for reliable classification. By generating randomized data partitions and varying feature subsets, the proposed methodology enables comprehensive evaluation of machine learning models under different experimental settings, thereby improving the robustness and consistency of the classification process.

Algorithm: Sequential Random Split with Random Feature Subsetting Selection

Input Parameters:

- D: Dataset with N Samples and M Features
- S: Number of sequential splits.
- $R = \{r_1, r_2, \dots, r_S\}$: Split ratios such that

$$\sum_{i=1}^S r_i = 1 \quad (1)$$

- $F = \{f_1, f_2, \dots, f_M\}$: List of features.
- $P = \{p_1, p_2, \dots, p_K\}$: Feature subset ratios such that

$$\sum_{j=1}^K p_j = 1 \quad (2)$$

Output: A list of m splits, each containing indices of data points and corresponding feature subsets.

Procedure:

Step-1: Data Preprocessing

- **Handle Missing Values:** Impute missing values using median imputation for numerical features and mode imputation for categorical features.
- **Normalization:** Normalize continuous features (e.g., RBC, HGB) to ensure uniform scaling.

$$X' = \frac{X - \mu}{\sigma} \quad (3)$$

- **Encoding Categorical Variables:** Convert categorical variables (e.g., Gender) using one-hot encoding.

Step-2: Verify Input Ratios

1. Check Split Ratios:
- Ensure the sum of Split_Ratios equals 1:

$$\sum Split_{ratios} = 1 \Rightarrow \sum_{i=1}^S r_i = 1 \quad (4)$$

- If not, raise an error: “Split ratios must sum to 1.”

Step-3: Calculate Split Sizes

- Determine the total number of observations,

$$Data_Len = |Data| \quad (5)$$

- Calculate the size of each split,

$$Split_Size_i = Split_Ratios_i \times Data_Len \quad (6)$$

- Store these in *split_sizes* to define proportions for training, validation and test sets.

$$n_i = N \cdot r_i \quad for \ i \in \{1, 2, \dots, S\} \quad (7)$$

Where n_i is the number of data points in the i^{th} split.

Step-4: Sequential Random Split

This step generates randomized data partitions while introducing controlled randomness to ensure reproducible and unbiased model evaluation.

1. Initialize the Lists: create an empty list *splits* to store data indices for each split.

2. Generate Indices:

List of indices from 0 to Data_Len-1

This step generates randomized data partitions while introducing controlled randomness to ensure reproducible and unbiased model evaluation.

1. Initialize the Lists: create an empty list *splits* to store data indices for each split.

2. Generate Indices:

List of indices from 0 to Data_Len-1

$$\begin{aligned} Indices &= [0, 1, 2, \dots, Data_Len - 1] \\ I &= \{0, 1, 2, \dots, N - 1\} \end{aligned} \quad (8)$$

Shuffle the indices using a random seed to ensure reproducibility:

$$Shuffled_indices = Shuffle(Indices, Seed = Shuffle_seed) \quad (9)$$

$$I = Shuffle(I) \quad (10)$$

3. Sequentially Divide Data: Assign Data Points to Splits

Starting with $start_idx = 0$, for each split size n_i

Calculate the end index:

$$end_idx = start_idx + n_i \quad (11)$$

For each $split_size_i$, select a sequential block of indices:

$$Split_i = Shuffled_Indices[start_idx:end_idx] \quad (12)$$

Assigning data points to the i^{th} split

$$\begin{aligned} S_i &= I[start_idx:end_idx] \\ start_idx &= end_idx \end{aligned} \quad (13)$$

Step-5: Feature Subset Selection

This step introduces feature variability to prevent overfitting

STEP-5.1: Validate Feature Subset Ratios

Check if the feature subset ratios sum to 1

$$\sum_{j=1}^K p_j = 1 \quad (14)$$

If not, raise an error: "Feature Subset ratios must sum to 1".

STEP-5.2: Calculate Subset Sizes: Determine the size of each feature subset using:

$$m_j = M \cdot p_j \text{ for } j \in \{1, 2, \dots, K\} \quad (15)$$

where m_j is the number of features in the j^{th} subset.

STEP-5.3: Random Feature Index Generation and Shuffling

Generate feature indices:

$$F_1 = \{0, 1, 2, \dots, M - 1\} \quad (16)$$

Randomly shuffle the feature indices:

$$F_1 = shuffle(F_1) \quad (17)$$

STEP-5.4: Assign Features to Subsets

Starting with $start_f_idx=0$, for each subset size m_j :

- Calculate the end index:

$$end_{f_idx} = start_f_idx + m_j \quad (18)$$

- Assign features to the j^{th} subset:

$$F_j = F_1[start_{f_idx}:end_{f_idx}] \quad (19)$$

- Update the start index:

$$start_f_idx = end_f_idx \quad (20)$$

Step-6: Attach Feature subsets to Data Splits

For each split S_i , pair the corresponding feature subset F_j with the data indices:

$$G_i = (S_i, F_j) \quad (21)$$

Where G_i represents the i th group containing data indices and the corresponding feature subset.

Step-7: Return Result

Return the list of groups

$$Output = \{G_1, G_2, \dots, G_s\} \quad (22)$$

End of Algorithm

The proposed framework should view as an integration of established data partitioning and Random Feature Subsetting strategies rather than as a fundamentally new learning algorithm. The primary objective of this study is to investigate whether combining Sequential Random Split (SRS) with Random Feature Subsetting improves the classification performance of imbalanced iron deficiency anemia datasets. This proposed framework does not introduce a new optimization objective, loss function, or adaptive learning mechanism. Instead, its contribution lies in the empirical evaluation of the integrated framework and its application to hematological data classification using ensemble learning models.

3.3. Research Flow

The overall workflow of the proposed framework illustrated in Fig.1. while section 3.2 presents the detailed methodological and algorithmic description, this section provides a high-level overview of the research process. The workflow begins with data collection and preprocessing, which includes data cleaning, managing missing values, and preparing the dataset for analysis. Feature engineering techniques applied to generate informative representations of the original hematological variables.

Following preprocessing, the dataset partitioned into training and testing subsets. To address class imbalance, the training data balanced using the Synthetic Minority Oversampling Technique (SMOTE). The balanced training data are subsequently processed using the Sequential Random Split (SRS) framework. Multiple feature subsets generated and used for model training and evaluation. The framework exposes machine learning models to diverse data partitions and feature combinations, enabling assessment across different experimental settings. Hyperparameter tuning is subsequently performed to optimize the performance of the model.

Lastly, the trained models evaluated on unseen data using accuracy, precision, recall, F1- Score and ROC-AUC metrics. This workflow provides a structures approach for assessing the effectiveness of the proposed SRS framework for iron deficiency anemia classification while ensuring systematic model development and evaluation.

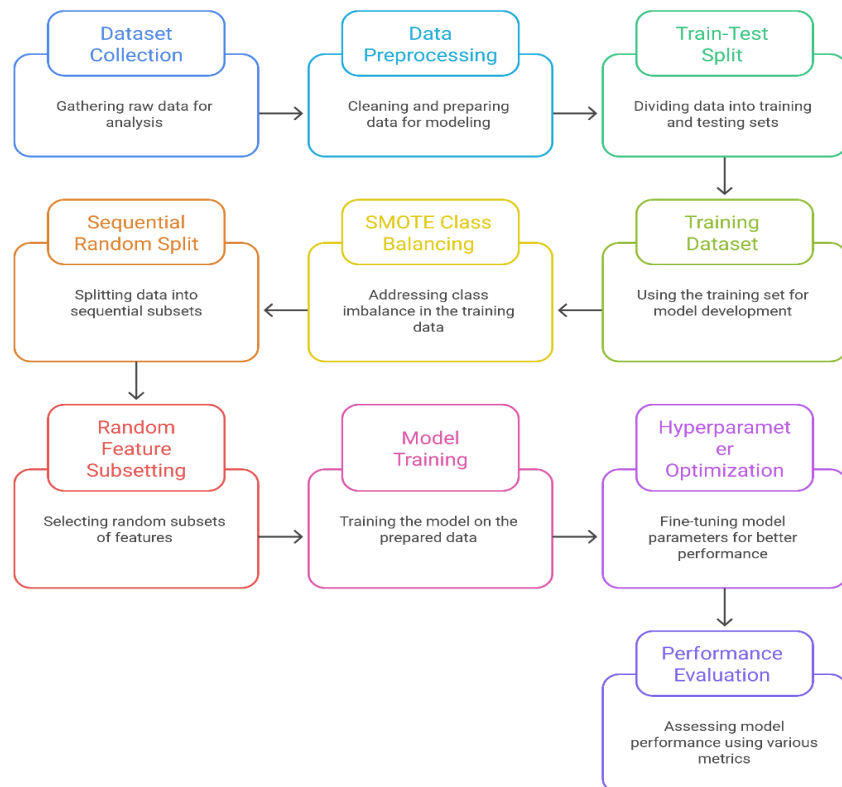


Fig. 1. Workflow of the proposed Sequential Random Split (SRS) framework for iron deficiency anemia classification

4. Experimentation and Results

To evaluate the effectiveness of the proposed Sequential Random Split (SRS) framework, experiments conducted using few ensemble learning methods, including Boosting, Bagging, Voting, Gradient Boosting, Extra Trees, Random Subspace, Random Patches, and Stacking. All classifiers evaluated using the same data partitioning strategy and performance metrics to ensure a fair and consistent comparison.

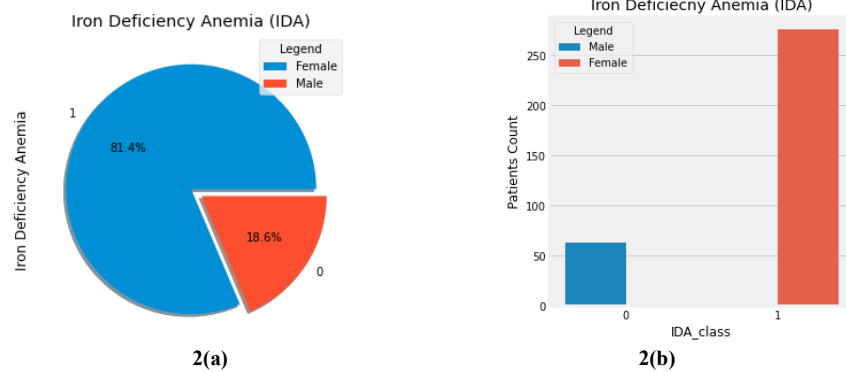


Fig. 2(a-b). Distribution of IDA and non-IDA cases in the study population.

The observed class imbalance highlights the necessity of employing robust sampling and classification strategies to prevent bias toward the majority class. Understanding class distribution is particularly important in medical diagnosis because minority-class misclassification may lead to missed identification of patients with iron deficiency anemia.

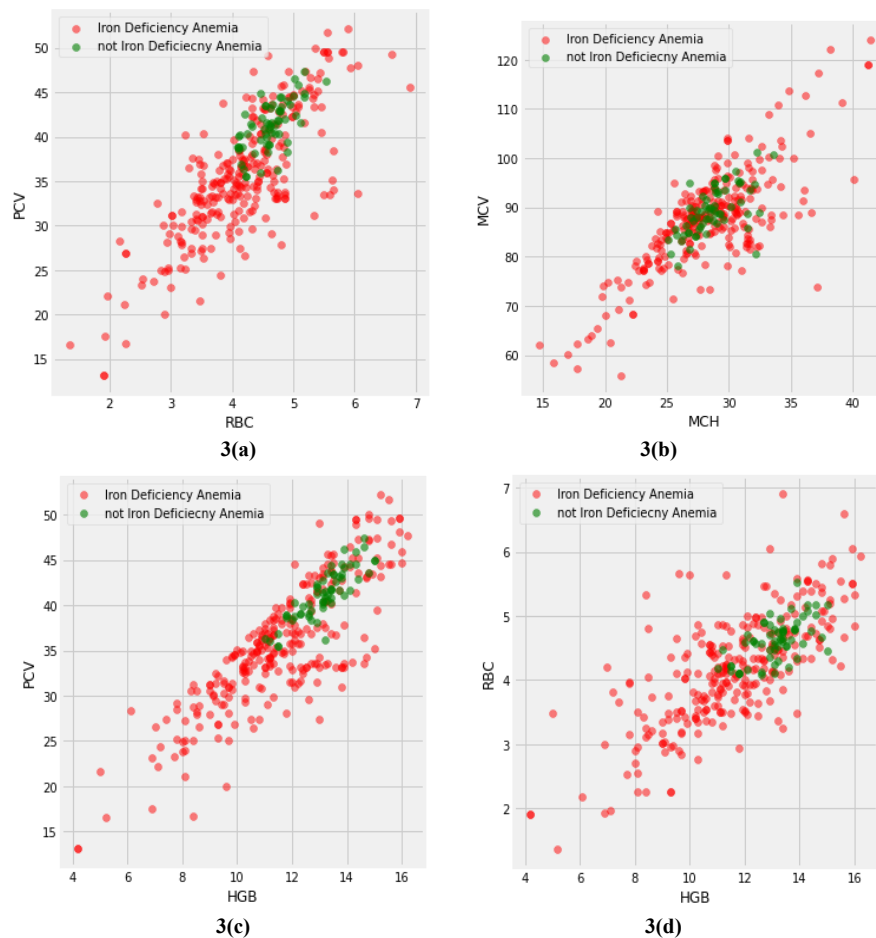


Fig. 3(a-d). Relationships among RBC, PCV, MCH, MCV, and HGB parameters.

Positive associations between these hematological indicators are consistent with known physiological relationships. The separation observed between IDA and non-IDA samples suggests that these parameters contribute valuable discriminatory information for machine learning-based anemia detection.

The scatter plots in Fig. 3(a), 3(b), and 3(c), 3(d) illustrate the following relationships:

- RBC vs. PCV: A positive correlation had expected, as higher red blood cell (RBC) counts increase the packed cell volume (PCV).
- MCH vs. MCV: This plot examines how changes in red blood cell volume (MCV) relate to hemoglobin content per cell (MCH), often showing a correlation.
- HGB vs. PCV: A strong positive correlation had anticipated between hemoglobin concentration (HGB) and PCV, as hemoglobin levels tied to the packed cell volume.
- HGB vs. RBC: This plot highlights the relationship between hemoglobin concentration and RBC count, helping to detect conditions like anemia.

These scatter plots offer a clear view of interrelationships between hematological parameters, allowing quick identification of trends and outliers for further analysis.

CATEGORICAL FEATURE: GENDER, RBC, AGE

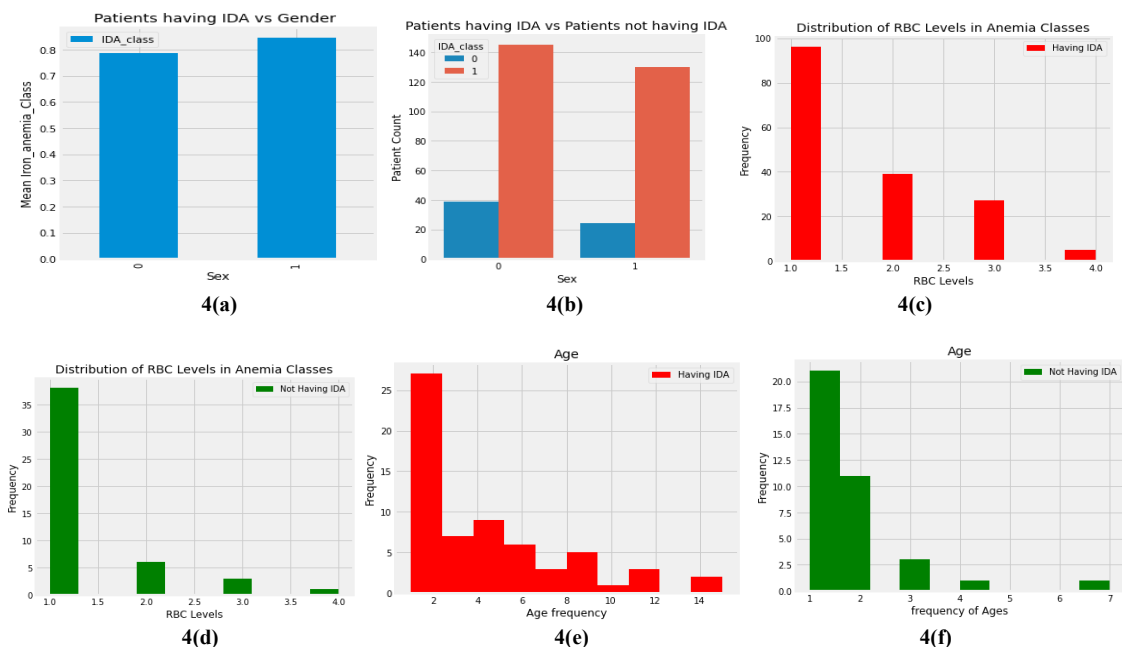
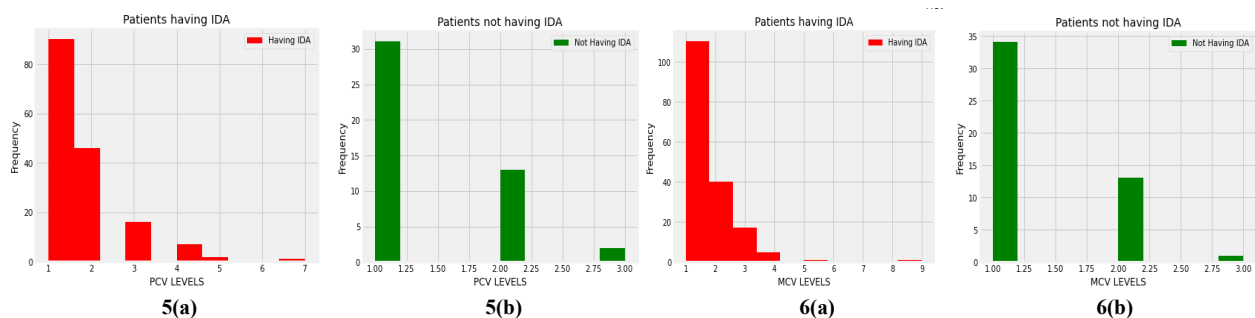


Fig. 4(a-f). Distribution of gender, RBC count, and age among IDA and non-IDA patients.

Variations in age and hematological characteristics indicate potential demographic and biological influences on anemia prevalence. These observations support the inclusion of demographic attributes alongside CBC parameters in predictive modelling.

These plots from Fig .4(a) to 4(f) collectively provide insights into the demographics and hematological characteristics of patients with and without IDA.

Continuous Features: PCV, MCV, MCH, MCHC, RDW, TLC, PLT, HGB



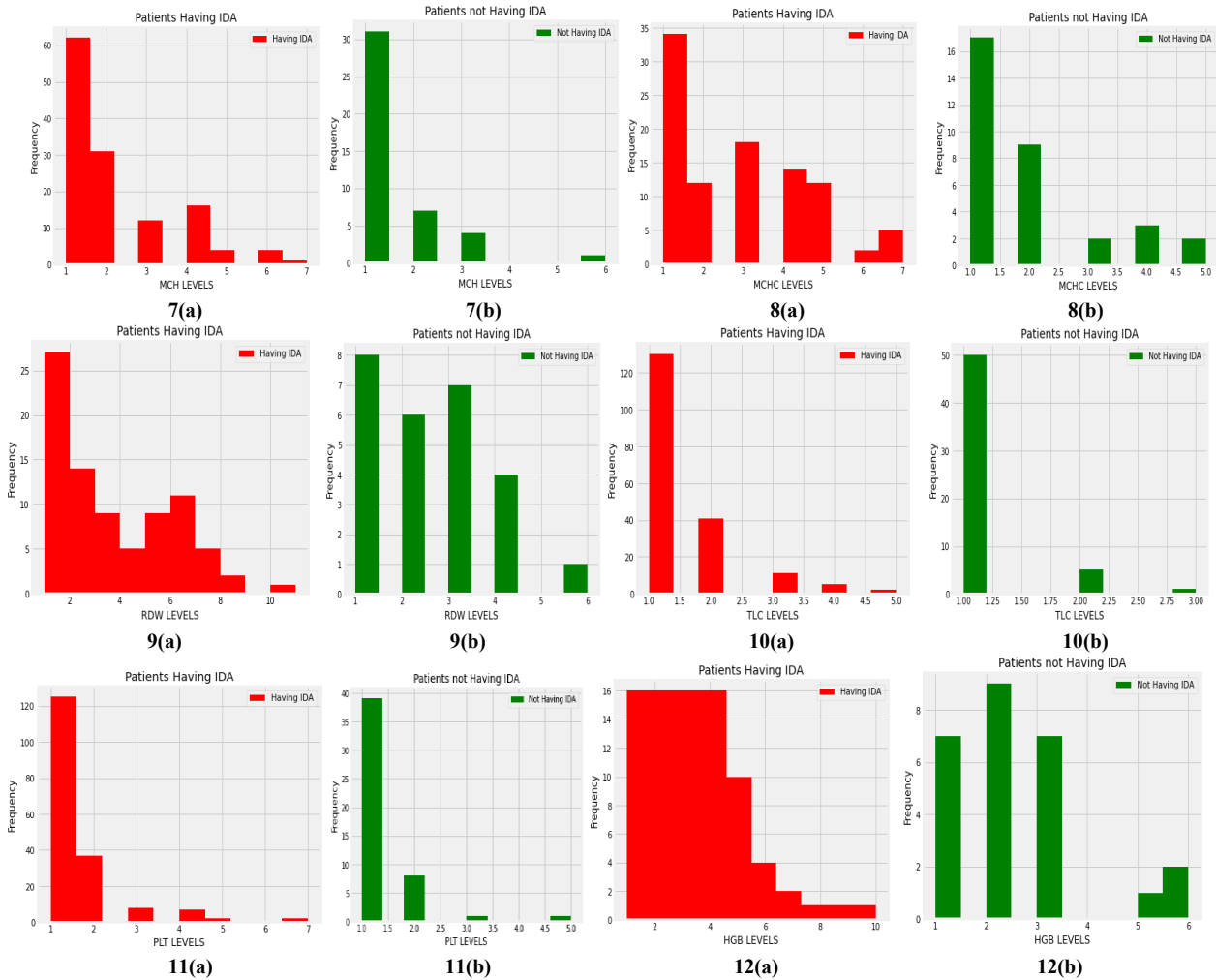


Fig. 5–12. illustrate the distributions of major hematological parameters among IDA and non-IDA patients.

Noticeable shifts in hemoglobin (HGB), packed cell volume (PCV), mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), and red cell distribution width (RDW) are consistent with established clinical manifestations of iron deficiency anemia. These differences indicate that CBC-derived features provide strong predictive information for automated IDA classification.

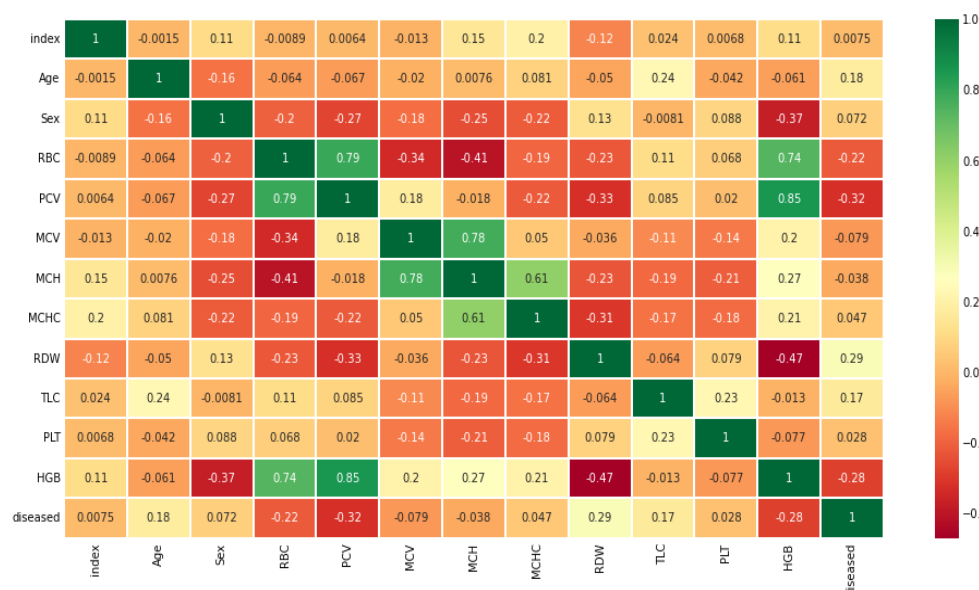


Fig. 13. Correlation matrix of hematological features.

Strong correlations among RBC, HGB, PCV, MCV, and MCH-related parameters reflect known physiological relationships within red blood cell production and function. The correlation analysis helps identify potentially redundant variables while highlighting informative feature groups for model development.

By visualizing the correlation matrix through a heatmap represented in Fig.13, we can efficiently identify strong relationships among parameters, informing further statistical analysis and hypothesis testing. This visualization enhances our understanding of interdependencies among haematological parameters and highlights key relationships for future investigation.

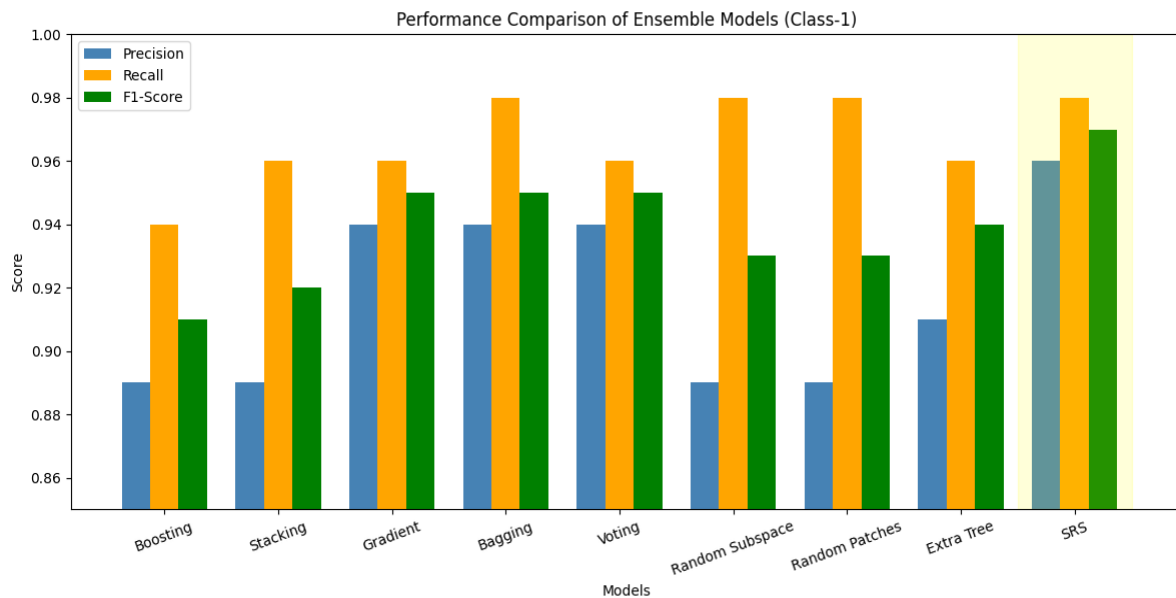


Fig. 14. Comparison of Precision, Recall, and F1-Score of ensemble learning models for Class-1 (Iron Deficiency Anemia).

The proposed Sequential Random Split (SRS) framework achieves the highest overall performance across the evaluated metrics compared with the baseline ensemble models.

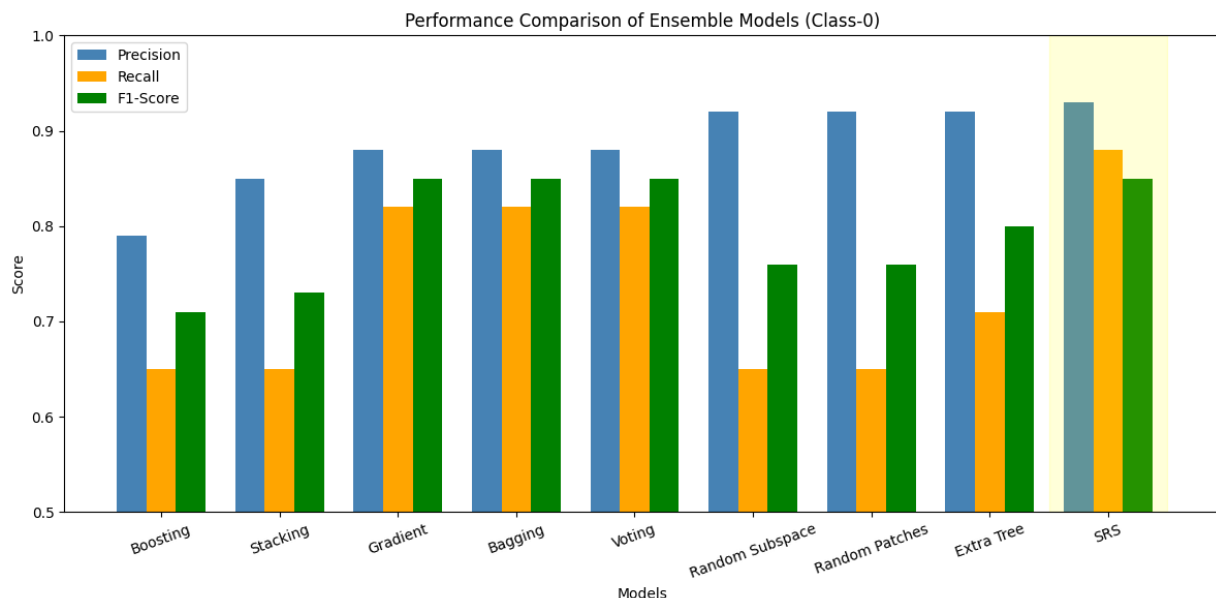


Fig. 15. Comparison of Precision, Recall, and F1-Score of ensemble learning models for Class-0 (Non-Iron Deficiency Anemia).

The proposed Sequential Random Split (SRS) framework demonstrates the highest overall performance among the evaluated ensemble models.

Figures 14 and 15 compare the Precision, Recall, and F1-Score achieved by the evaluated ensemble learning models for Class-0 (Non-Iron Deficiency Anemia) and Class-1 (Iron Deficiency Anemia), respectively. The proposed Sequential Random Split (SRS) framework consistently achieved the highest overall performance across the three-evaluation metrics.

For Class-0, SRS obtained the highest Precision (0.93), Recall (0.88), and F1-Score (0.85), indicating improved recognition of the majority class. Similarly, for Class-1, SRS achieved the highest Precision (0.96), Recall (0.98), and F1-Score (0.97), demonstrating superior capability in identifying iron deficiency anemia cases compared with the other ensemble learning models. These results indicate that integrating Sequential Random Split with Random Feature Subsetting provides a more effective learning strategy for the evaluated imbalanced dataset.

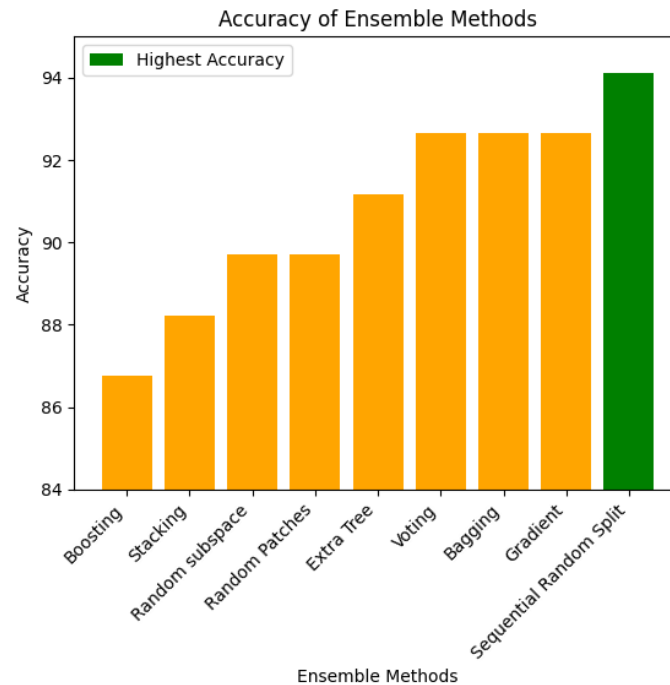


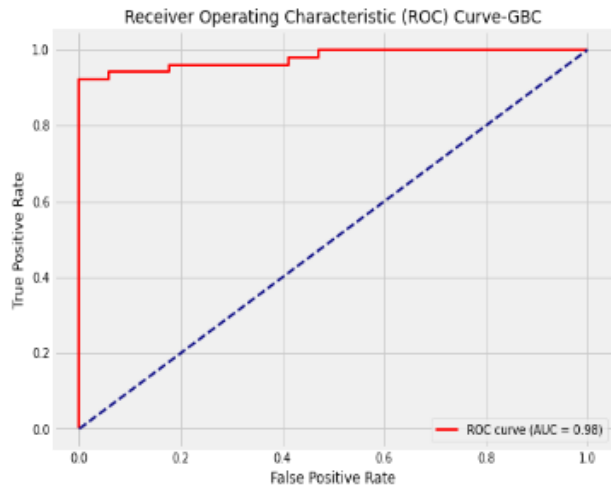
Fig. 16. Accuracy comparison among ensemble learning models and the proposed SRS framework.

The plot shown in Fig.16 visualizes the accuracy achieved by various ensemble methods and our proposed model. Each bar represents a different ensemble method and our proposed model, and the x-axis lists the ensemble methods and proposed model evaluated. The proposed Sequential Random Split (SRS) framework was evaluated using ensemble learning classifiers, including Boosting, Bagging, Voting, Gradient Boosting, Extra Trees, Random Subspace, Random Patches, and Stacking. The same experimental protocol, training/testing partitions, and evaluation metrics were applied to all models to ensure a consistent comparison.

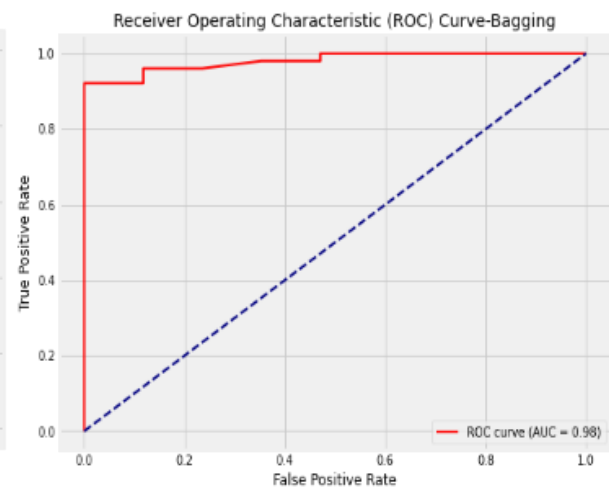
Table 2. Comparison of different Ensemble Models and Proposed Model

Ensemble Model Names	Accuracy
Boosting	86.76
Stacking	88.23
Random Subspace	89.70
Random Patches	89.70
Extra Tree	91.17
Voting	92.64
Bagging	92.64
Gradient	92.64
Sequential Random Split	94.11

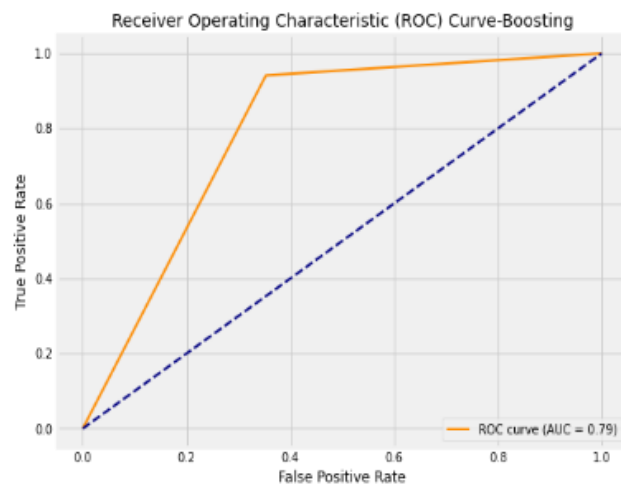
The Sequential Random Split (SRS) model achieved the highest accuracy among the ensemble models at 94.11%. Other strong performers included Bagging, Voting, and Gradient models, all at 92.64%. The results emphasize the effectiveness of the SRS model in enhancing predictive performance compared to the other ensemble methods are in above table 2.



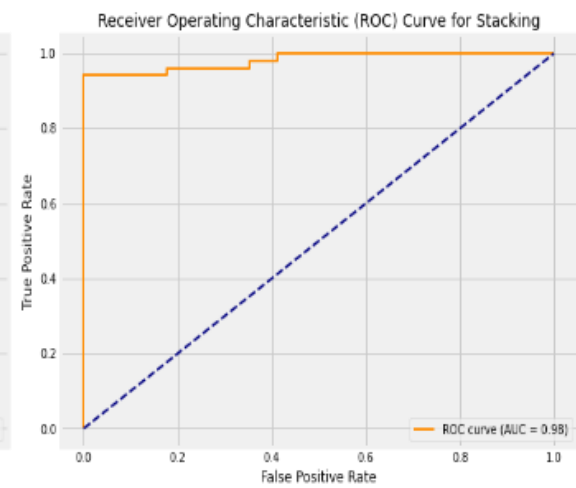
17(a)



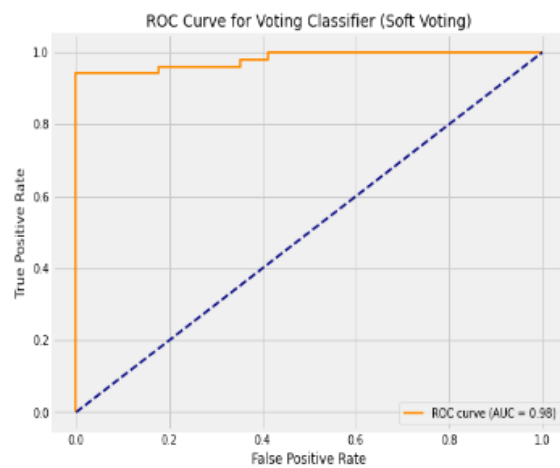
17(b)



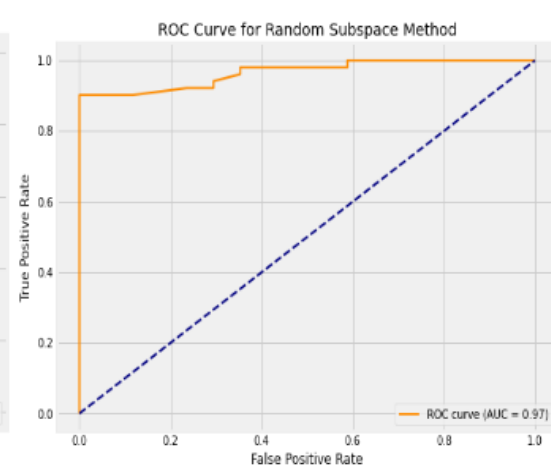
17(c)



17(d)



17(e)



17(f)

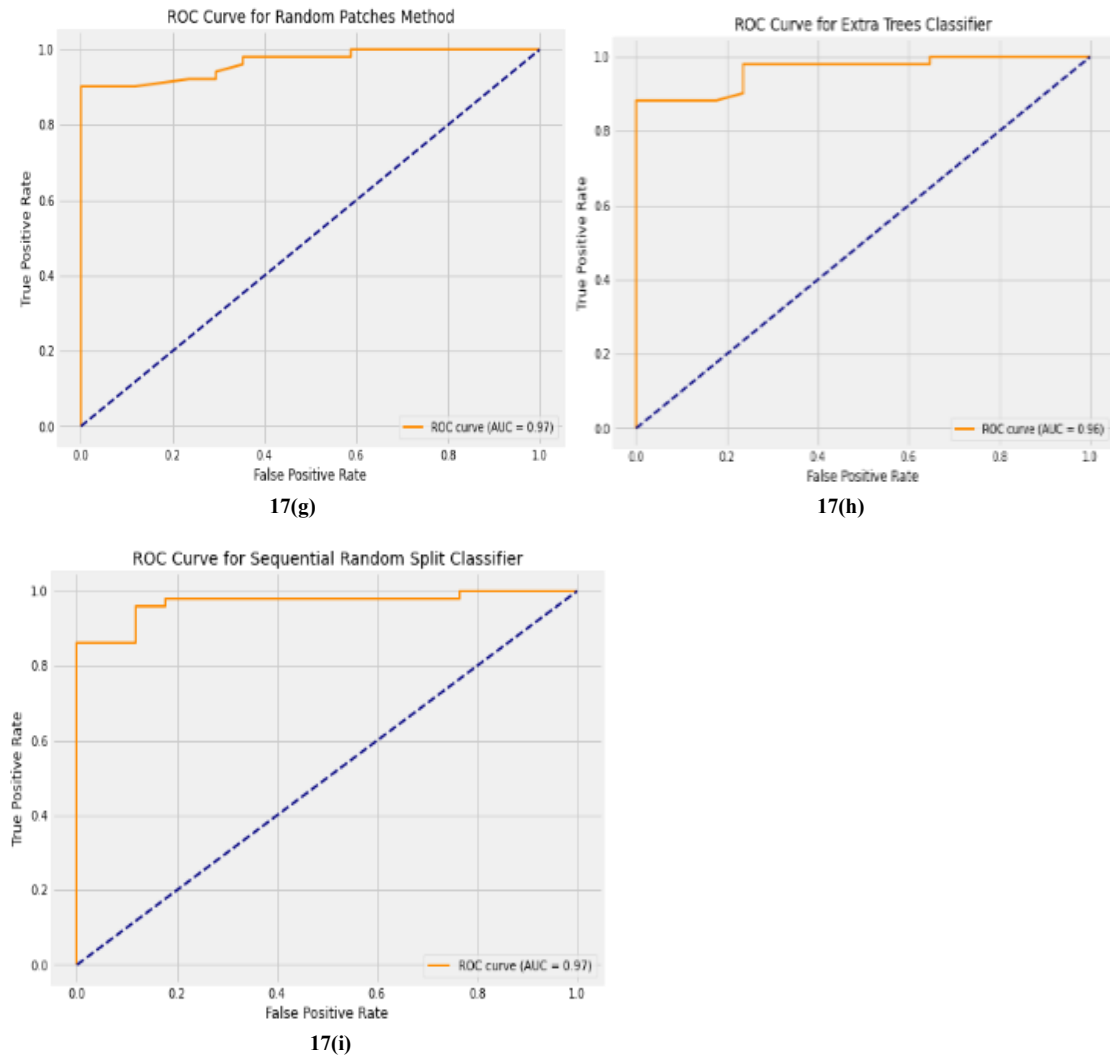


Fig. 17(a-i). ROC curves for the evaluated classifiers.

Fig.17(a) Gradient Boosting Classifier: Robust performance with a high AUC, indicating effective discrimination. Fig.17(b) Bagging Classifier: Robust and stable predictions, reflected in its favourable AUC. Fig.17(c) Boosting Classifier: Sharp rise in sensitivity, demonstrating its effectiveness in identifying positive cases. Fig.17(d) Stacking Classifier: Combines multiple models for improved predictive power and classification performance. Fig.17(e) Voting Classifier: Aggregates predictions with balanced sensitivity and specificity. Fig.17(f) Random Subspace Classifier: Solid performance by leveraging feature subsets. Fig.17(g) Random Patches Classifier: Effectively manages data variability with reliable predictions. Fig.17(h) Extra Tree Classifier: Strong capture of complex data patterns. Fig.17(i) Sequential Random Split Classifier: Promising results, effectively addressing dataset imbalance.

All models demonstrate discriminative capability, while the proposed SRS framework achieves competitive ROC performance. The ROC analysis confirms the ability of the proposed approach to distinguish between IDA and non-IDA cases across a range of classification thresholds.

5. Conclusion

This study proposes an integrated machine learning framework Sequential Random Split framework combined with Random feature Subsetting for detecting iron deficiency anemia (IDA) using a sequential random split approach on an imbalanced dataset. By employing a structured data preparation process that includes imputation for missing values, categorical feature encoding, and advanced feature engineering, the model effectively utilizes informative feature representations derived from hematological parameters. The use of Sequential Random Split combined with Random Feature Subsetting selection enhances generalization by exposing the model to diverse training subsets, reducing overfitting risks, and ensuring robust predictive performance. Additionally, the dual application of Sequential Random Split, along with resampling techniques like SMOTE, effectively addresses class imbalance, enabling balanced learning from both majority and minority classes.

The experimental results suggest that the proposed framework may be suitable for real-world medical diagnostics due to its adaptability and comprehensive evaluation metrics, including precision, recall, F-score, and AUC, which provide a balanced assessment of model accuracy. Its ability to manage imbalanced datasets ensures reliable predictions even in rare disease cases, making it particularly useful for clinical diagnostics.

Despite the encouraging classification results, still few limitations should acknowledge. The proposed framework represents an integration of established data partitioning and Random Feature Subsetting strategies rather than a new optimization-based learning model. Furthermore, the study evaluated using a single publicly available dataset, which may limit the generalizability of the findings to other clinical populations. In addition, the performance of the proposed framework under extremely imbalanced class distributions or across heterogeneous clinical populations not investigated in the present study and should explore in future research.

The experimental evaluation focused on established ensemble learning methods and did not include comparisons with more recent state-of-the-art machine learning models. In addition, the study emphasizes predictive performance and does not include component-wise ablation analysis, confidence interval estimation, or repeated experimental evaluations to quantify the contribution and statistical variability of each framework component. Future work will focus on validating the proposed framework using larger multicentre datasets, investigating optimization-driven feature selection and adaptive sampling strategies, incorporating recent machine learning models, and performing comprehensive statistical analyses to further evaluate the robustness and generalizability of the proposed framework.

Additionally, expanding the dataset to include diverse patient demographics and evaluating the model in different clinical settings would improve generalizability. Investigating other resampling techniques or hybrid models that combine synthetic data generation with sequential random splitting could further refine the model's performance. Overall, these findings suggest that the proposed framework may be useful for healthcare-oriented machine learning applications and warrant further validation on larger and more diverse datasets.

All the Declarations and Statements

Author Contributions Statement

S. N. Lakshmi Malluvalasa- Conceptualization, Investigation, Formal Analysis, Data Curation, Writing – Original Draft, Writing – Review & Editing.
Sajana T. - Methodology, Supervision, Writing – Review & Editing.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare that they have no known competing financial interests.

Funding Declaration

This research received no external funding.

Data Availability Statement

The dataset used in this study is publicly available in the Mendeley Data repository and can be accessed at Mendeley Data – Dataset for Anemia Classification Based on Hematological Parameters

Ethical Declarations

The study utilized a publicly available, anonymized dataset from the Mendeley Data repository (Reference [37]). The authors did not collect any data directly from human participants, and no identifiable personal information was accessed. Consequently, ethical approval and informed consent were not required for this research.

Acknowledgments

The authors express their sincere gratitude to the research supervisor, Dr. T. Sajana, for valuable methodological guidance, constructive suggestions, and continuous academic support throughout this research. The first author also acknowledges the support and facilities provided by the affiliated institution for conducting this research.

Declaration of Generative AI in Scholarly Writing

The authors used generative artificial intelligence tools solely only to improve the language, grammar, and readability of the manuscript. All scientific content, analysis, and conclusions are the original work of the authors, who take full responsibility for the final manuscript.

Abbreviations

The following abbreviations are used in this manuscript:

AI-Artificial Intelligence
AUC-Area Under the Curve
CBC-Complete Blood Count
CNN-Convolutional Neural Network
F1-Score-Harmonic Mean of Precision and Recall
GB-Gradient Boosting
HGB-Hemoglobin
IDA-Iron Deficiency Anemia
ML-Machine Learning
MCH-Mean Corpuscular Hemoglobin
MCHC-Mean Corpuscular Hemoglobin Concentration
MCV-Mean Corpuscular Volume
PCV-Packed Cell Volume
PLT-Platelet Count
RBC-Red Blood Cell Count
RDW-Red Cell Distribution Width
RFE-Recursive Feature Elimination
ROC-Receiver Operating Characteristic
RNN-Recurrent Neural Network
RSS-Random Subset Selection
SRS-Sequential Random Split
SMOTE-Synthetic Minority Over-sampling Technique
SVM-Support Vector Machine
TLC-Total Leukocyte Count
WHO-World Health Organization
XAI-Explainable Artificial Intelligence

Appendix A/B/C..., with appendix title

N/A

References

- [1] K. N. Kawo, Z. G. Asfaw, and N. Yohannes, "Multilevel analysis of determinants of anemia prevalence among children aged 6–59 months in Ethiopia: Classical and Bayesian approaches," *Anemia*, vol. 2018, Art. no. 3087354, 2018.
- [2] J. E. Feusier et al., "Large-scale identification of clonal hematopoiesis and mutations recurrent in blood cancers," *Blood Cancer Discovery*, vol. 2, no. 3, pp. 226–237, 2021, Doi: 10.1158/2643-3230.BCD-20-0223.
- [3] World Health Organization, Hemoglobin Concentrations for the Diagnosis of Anemia and Assessment of Severity. Geneva, Switzerland: WHO, 2011. WHO/NMH/NHD/MNM/11.1.
- [4] A. Baruah and S. Gautam, "Prevalence and predictors of iron deficiency anemia in adolescent girls in India," *International Journal of Community Medicine and Public Health*, vol. 10, no. 10, pp. 3917–3923, Oct. 2023.
- [5] T. K. Yıldız, N. Yurtay, and B. Öneç, "Classifying anemia types using artificial learning methods," *Engineering Science and Technology*, an International Journal, vol. 25, Art. no. 100996, 2021.
- [6] World Health Organization, "Anemia," *WHO Fact Sheet*, 2023. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/anaemia>. [Accessed: Oct. 3, 2023].
- [7] S. A. Ali, U. S. Khan, and A. S. Feroz, "Prevalence and determinants of anemia among women of reproductive age in developing countries," *Journal of the College of Physicians and Surgeons Pakistan*, 2020 Feb;30(2):177-186. DOI: 10.29271/jcpsp.2020.02.177.
- [8] 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, 2022.
- [9] R. Lassila and J. W. Weisel, "Role of red blood cells in clinically relevant bleeding tendencies and complications," *Journal of Thrombosis and Haemostasis*, vol. 21, no. 6, pp. 1415–1425, 2023.
- [10] M. D. Cappellini, K. M. Musallam, and A. T. Taher, "Iron deficiency anemia revisited," *Journal of Internal Medicine*, vol. 287, no. 2, pp. 153–170, 2020.
- [11] R. S. Hussien, S. I. A. Jabuk, Z. M. Altaee, and A. M. K. Al-Maamori, "Review of anemia: Types and causes," *European Journal of Research Development and Sustainability (EJRDS)*, Vol. 4 No 07, July 2023, ISSN: 2660-5570,
- [12] D. T. Lee and M. L. Plesa, "Anemia," in *Family Medicine: Principles and Practice*. Cham, Switzerland: Springer, 2022, pp. 1815–1829.
- [13] A. Al-Naseem, A. Sallam, S. Choudhury, and J. Thachil, "Iron deficiency without anemia: A diagnosis that matters," *Clinical Medicine*, vol. 21, no. 2, pp. 107–113, 2021.
- [14] B. J. Bain, *Blood Cells: A Practical Guide*, 6th ed. Hoboken, NJ, USA: Wiley-Blackwell, 2021.

- [15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [16] Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, nos. 1–3, pp. 389–422, 2002.
- [17] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.
- [18] E. Tuv, A. Borisov, G. Runger, and K. Torkkola, "Feature selection with ensembles, artificial variables, and redundancy elimination," *Journal of Machine Learning Research*, vol. 10, pp. 1341–1366, 2009.
- [19] S. Sharma, V. Khullar, and A. Luhach, "Comparative study of back-propagation and PSO based back-propagation for anemia diagnosis in pregnant ladies," *International Journal of Information Technology*, vol. 9, no. 2, pp. 104–110, 2017.
- [20] M. Gupta, M. O'Halloran, and A. Gupta, "Investigation of anemia and the dielectric properties of human blood at microwave frequencies," *IEEE Access*, vol. 6, pp. 3536–3547, 2018.
- [21] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [22] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [23] L. Liu, S. Tan, Y. Li, J. Luo, and W. Zhang, "An early aortic dissection screening model and applied research based on ensemble learning," *Journal of Translational Medicine*, vol. 18, Art. no. 18, 2020.
- [24] B. Schölkopf, A. J. Smola, and K.-R. Müller, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.
- [25] X. Liu, L. Wang, and L. Zhang, "Discovering the structure of large-scale networks," *Journal of Complex Networks*, vol. 16, no. 4, pp. 689–704, 2008.
- [26] 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.
- [27] B. Ç. Yavuz, T. K. Yıldız, N. Yurtay, and Z. Pamuk, "Comparison of K nearest neighbors and regression tree classifiers used with clonal selection algorithm to diagnose hematological diseases," *AJIT-e: Online Academic Journal of Information Technology*, vol. 5, no. 16, pp. 7–20, 2014.
- [28] S. A. Sanap, M. Nagori, and V. Kshirsagar, "Classification of anemia using data mining techniques," in *SEMCCO: International Conference on Swarm, Evolutionary, and Memetic Computing*, Springer, 2011, pp. 113–124.
- [29] N. Amin and A. Habib, "Comparison of different classification techniques using WEKA for hematological data," *American Journal of Engineering Research*, vol. 4, no. 3, pp. 55–61, 2015.
- [30] A. R. Khawaga, E. A. Shehab, and S. A. Shehab, "Anemia diagnosis and prediction based on machine learning," *Kafrelsheikh Journal of Information Sciences*, vol. 4, no. 2, pp. 1–9, Nov. 2023.
- [31] J. R. Khan et al., "Machine learning algorithms to predict the childhood anemia in Bangladesh," *Journal of Data Science*, vol. 17, no. 1, pp. 195–218, 2019.
- [32] M. Jaiswal, A. Srivastava, and T. J. Siddiqui, "Machine learning algorithms for anemia disease prediction," in *Recent Trends in Communication, Computing, and Electronics*, Singapore: Springer, 2019, pp. 275–285.
- [33] E. M. T. El-Kenawy, "A machine learning model for hemoglobin estimation and anemia classification," *International Journal of Computer Science and Information Security*, vol. 17, no. 1, pp. 100–108, 2019.
- [34] A. Dixit et al., "Prediction of anemia disease using machine learning algorithms," in *Intelligent Computing and Networking (IC-ICN 2022)*, Singapore: Springer, 2023, pp. 229–238.
- [35] S. Dogan and I. Turkoglu, "Iron deficiency anemia detection from hematology parameters by using decision tree," *International Journal of Science and Technology*, vol. 3, no. 1, pp. 85–92, 2008.
- [36] M. F. Young, B. M. Oaks, S. Tandon, R. Martorell, K. G. Dewey, and A. Wendt, "Maternal hemoglobin concentrations across pregnancy and maternal and child health: A systematic review and meta-analysis," *Annals of the New York Academy of Sciences*, vol. 1450, no. 1, pp. 47–68, 2019.
- [37] Mendeley Data, "Dataset for Anemia Classification Based on Hematological Parameters," ver. 1. [Online]. Available: <https://data.mendeley.com/datasets/dy9mfjchm7/1>.
- [38] P. Appiahene, J. W. Asare, E. T. Donkoh, G. Dimauro, and R. Maglietta, "Detection of iron deficiency anemia by medical images: A comparative study of machine learning algorithms," *BioData Mining*, vol. 16, Art. no. 2, 2023, Doi: 10.1186/s13040-023-00319-z.
- [39] B. S. D. Darshan, N. Sampathila, G. M. Bairy, et al., "Differential diagnosis of iron deficiency anemia from aplastic anemia using machine learning and explainable artificial intelligence utilizing blood attributes," *Scientific Reports*, vol. 15, Art. no. 505, 2025, Doi: 10.1038/s41598-024-84120-w.
- [40] W. Tepakhan, W. Srisintorn, T. Penglong, et al., "Machine learning approach for differentiating iron deficiency anemia and thalassemia using Random Forest and Gradient Boosting algorithms," *Scientific Reports*, vol. 15, Art. no. 16917, 2025, doi: 10.1038/s41598-025-01458-5.

Authors' Profiles



S. N. Lakshmi Malluvalasa, M.Tech, an Assistant Professor in the Department of Computer Science and Engineering at Koneru Lakshmaiah Education Foundation, Vaddeswaram, India. Her research interests include Artificial Intelligence, Machine Learning, and Deep Learning. She has contributed to academic research through publications in reputed journals and conferences and holds three patents in emerging areas of technology. Her work focuses on the development and application of intelligent systems, data-driven solutions, and advanced computational techniques. She is actively engaged in teaching, research, and mentoring students in innovative domains of computer science.



Dr.T. Sajana received the B.Tech degree from Acharya Nagarjuna University Andhra Pradesh, India, the M. Tech degree from the JNT University Kakinada, Andhra Pradesh and the Ph.D. degree from the KL University Vaddeswaram), Andhra Pradesh. She is currently an Associate Professor in the Department of Computer Science and Engineering at KL University. Her research interests are in Machine learning, Deep Learning and Data Analytics. She is a senior member of the CSI. Her research interests include Artificial Intelligence, Machine Learning, and Deep Learning. She is a recognized Ph.D. supervisor and actively mentors doctoral scholars while contributing to research through publications in reputed journals and conferences. Her academic interests focus on intelligent computing, data-driven technologies, and innovative AI-based solutions.

How to cite this paper

S. N. Lakshmi Malluvalasa, T. Sajana, "Sequential Random Split with Random Feature Subsetting for imbalanced Iron Deficiency Anemia Classification: A Machine Learning Study", International Journal of Engineering and Manufacturing (IJEM), Vol.16, No.4, pp.40-57, 2026. DOI:10.5815/ijem.2026.04.03