A Comparative Study of Anemia Classification Algorithms for International and Newly CBC Datasets

https://doi.org/10.3991/ijoe.v19i06.38157

Safa S. Abdul-Jabbar[1,2](✉), Alaa K. Farhan[2], Alexander S. Luchinin[3]
[1]Computer Science Department, College of Science for Women, University of Baghdad, Baghdad, Iraq
[2]Computer Science Department, University of Technology, Baghdad, Iraq
[3]Federal State Budgetary Institution of Science, Kirov Research Institute of Hematology and Blood Transfusion of the Federal Medical and Biological Agency, Kirov, Russia
safa.s@csw.uobaghdad.edu.iq

Abstract—The healthcare field has experienced a significant increase in data generation due to the emergence of modern applications and the internet. Consequently, understanding and extracting meaningful information from these extensive datasets has become a critical factor in the success of any application in this sector. Digital analytics and classification tools can assist in handling the challenges of processing large datasets to produce highly consistent, logical, and information-rich summaries. This paper presents several analytics methodologies based on literature that can be used as pre-processing steps to determine dataset characteristics. The study conducted a comparative analysis of twelve classification algorithms using two international datasets to measure their efficiency accurately. The outcome of the analysis step will assist researchers in selecting the most suitable algorithm for each dataset's characteristics, resulting in more organized and thorough results. The study revealed that four algorithms, namely Logitboost, Random Forest, XGBoost, and Multilayer Perceptron, achieved the best accuracy. The XGBoost algorithm, which produced the highest accuracy, was used to classify new CBC datasets collected from various hospitals in Iraq for Hematology studies and statistics. Future research should investigate combining algorithms to leverage their benefits while overcoming their limitations. Overall, using digital analytics tools and algorithms in healthcare can provide critical insights into large datasets, leading to improved disease diagnosis outcomes and the advancement of medical knowledge.

Keywords—anemia diagnosis model, data analytics tools, analytics methodologies, hematology, CBC dataset

1 Introduction

Over the past two decades, artificial intelligence (AI) has seen extraordinary progress and value expansion, along with its successful introduction for resolving challenging data-related tasks [1] [2]. Recent studies have demonstrated the potential of AI in predicting and diagnosing various health conditions, including anemia.

Anemia is a common condition characterized by a deficiency of red blood cells or hemoglobin in the blood, leading to fatigue, weakness, and other health problems [3] [4]. On the other hand, anemia is the most common disease in many countries, such as Pakistan, India, Iran, and Iraq. AI-based models have been developed to aid in diagnosis and its underlying causes. For example, a study by Lippi et al. used machine learning algorithms to analyze routine blood tests and accurately predict the presence of anemia with high sensitivity and specificity [5]. Another study by Li et al. developed an AI-based model to predict the risk of iron deficiency anemia in pregnant women, which could aid in early intervention and prevention of the condition [6].

Moreover, AI-based models can also aid in diagnosing anemia's underlying causes, such as chronic kidney disease. A study by Yan et al. demonstrated that an AI-based model could accurately predict the risk of anemia in patients with chronic kidney disease, which could help in diagnosing and treating the condition [7].

These examples demonstrate the potential of AI in diagnosing and predicting anemia, which could lead to more accurate and efficient diagnoses and treatments, improving patient outcomes. Therefore, many research papers were published to identify and classify the types of anemia. For example, in 2017, Kandhro et al. published a research paper to differentiate between Thalassemia traits (TTs) and iron deficiency anemia (IDA). They proposed a new formula for determining the cut-off value to differentiate between TTs and IDA. The proposed formula used by Random forest and Decision Tree algorithms show 100% classification accuracy [3]. While in 2019, Jaiswal et al. investigated several Machine Learning (ML) algorithms with 200 Complete Blood Count (CBC) test samples with 18 features. The results show a maximum accuracy of 96.09 by the Naive-Bayes Algorithm [4]. In the same year, Alsheref and Gomaa published a research paper that provided a comparative study to evaluate the performance of the ML Algorithms on a pathological dataset that contained 668 records. The results show that the LogitBoost algorithm reached an accuracy level of 98.16% [8]. While diagnosing 1577 individuals with haematological neoplasms, the usual CBC parameters and research CBC items from a haematology analyzer were gathered in 2022 by Haider et al. They leverage the hidden trend by improving the auguring accuracy of these prospective morphometric parameters in the differentiation leukemias. CBC parameter-driven artificial neural network (ANN) prediction modelling was created. The results show that the maximum practice result of using the proposed model for training was 83.1%, and testing was 89.4.7% [9]. Also, in the same year, Vohra et al. tested multi-classification algorithms using the Complete blood count test for 400 patient samples applying botten10 cross-validation and hold-out strategies to determine the type of anemia. The maximum accuracy for all proposed systems was 94.44% [10].

This paper provides several analysis tools as a preprocessing step. Also, a comparative study of several methods used in recent years illustrates the advantage and disadvantages of each one. Finally, provide an open-source dataset for anemia diseases using CBC tests, then apply the best methods on several labelled datasets to test their efficiency and select the best one that can be used to classify these new datasets to diagnose anemia classes.

2 Research methodologies

In this paper, a ML Model is being designed that helps doctors and other health sector researchers. The main steps in the proposed Model, as illustrated in Figure 1, can explain as follows:

1. Preprocessing Step (Prescriptive Analytics Operations).
2. Learning Step (using different classifiers).
3. Evaluate the Anemia Classification Model Step.
4. Data Collection and Labeling Step.

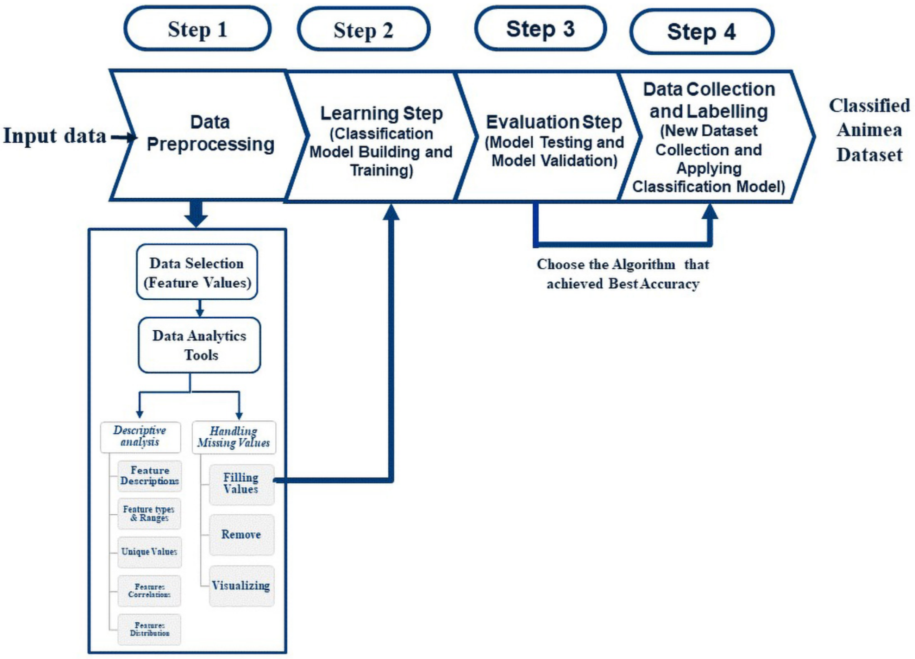

Step 1
Step 2
Step 3
Step 4
Input data
Data Preprocessing
Learning Step (Classification Model Building and Training)
Evaluation Step (Model Testing and Model Validation)
Data Collection and Labelling (New Dataset Collection and Applying Classification Model)
Classified Animea Dataset
Choose the Algorithm that achieved Best Accuracy
Data Selection (Feature Values)
Data Analytics Tools
Descriptive analysis
Feature Descriptions
Feature types & Ranges
Unique Values
Features Correlations
Features Distribution
Handling Missing Values
Filling Values
Remove
Visualizing

Fig. 1. General diagram for the proposed system

All these steps will be described in the following sections:

2.1 Preprocessing step

The input of this step is the complete dataset features as an excel file, using the Prescriptive Analytics Operations that can provide the following tools:

1. Feature Descriptions (using Razy Algorithm for words matching operation [11]).
2. Feature type and values.
3. Feature Histogram and statistical information for each one.
4. Checking missing values also plots the percentage for all the features.

5. Find the Unique values.
6. Features Distribution.
7. Features Correlations.
8. Drop or Filling Missing Features.

In this step, we take an image or PDF file as input and use it to extract valuable information that can be inserted into an Excel file. The extraction process is carried out through a matching operation, where the algorithm extracts the values of each feature from the image or PDF file. To perform this matching operation, we employed the RAZY algorithm, which has proven to be effective in extracting accurate information to fill in the feature values in the Excel file [11]. This approach can efficiently convert unstructured information into structured data that can be analyzed and utilized for various purposes. While the output is a visual description of the overall data characteristics used to describe the nature of the collected data. In this paper, several tools are used for visualizing the data characteristics, such as:

- Feature Description, Feature Types and Normal Values Range:
  These tools provide descriptive information about all features in the Dataset, what it means, and the typical ranges or meanings for each value. For example: What are the typical ranges of each feature? What is the feature data type? How much space is used to store each value? What is the maximum/minimum value of each feature? …, etc.
- Find Unique Values:
  This tool is a valuable help to researchers and medical professionals, allowing them to identify and determine outlier values within their data. By utilizing the tool, users can quickly and accurately analyze their data to determine whether the outlier values are simply noise or indicative of a specific case that requires further investigation. With the ability to accurately distinguish between noise and relevant information, researchers and medical professionals can make informed decisions regarding the direction of their research or the appropriate course of action to take in the case of an outlier value. This tool provides a critical resource for accurate and meaningful data analysis, ultimately leading to improved research outcomes and patient care.
- Feature Histogram and Statistical Information:
  The technique employed in this tool is simple and highly practical, providing a powerful means of summarizing and expressing data concisely and meaningfully. The data for each feature is represented as a vector through the technique, with each element reflecting the frequency of its appearance within the data. By utilizing this method, we can gain a deeper understanding of the underlying patterns and trends within the data, ultimately leading to more effective data analysis and interpretation. Moreover, the simplicity of the technique ensures that it is easily implementable and accessible, making it an invaluable tool for researchers, analysts, and other data professionals seeking to derive insights from their data. The resulting vectors are highly informative, providing a clear and concise means of representing data that is both intuitive and highly actionable.
- Drop or Filling the Missing Value:
  This is an essential tool when dealing with Machine Learning (ML) or Deep learning (DL) because we need to train and learn the model with as much data as possible with accurate data to keep the result at an acceptable level of accuracy. So, in this

paper, two types of handling the missing data were provided. First, missing values are deleted if the volume of data is small compared to the total volume of data. Second, if the missing values are relatively large compared to the total size, we fill these values with the median value of each feature.

- Feature Correlations, Feature Distribution, and Missing Data Information:
  These tools provide an informative summary for doctors or programmers about the feature correlation and distribution. This is because the potential applications of ML models in healthcare are vast and can be used for anemia disease classification that doctors can use. Also, it can be used by programmers by providing preliminary information about the benefits and disadvantages of each algorithm and determining the best algorithm for use in this field.

2.2 Learning step

To design and implement a new hybrid model that can be used to deal with the CBC data. We first provide a Comparative study of traditional ML Algorithms. In this paper, many different algorithms were explored to decide which is the best for dealing with Pathological reports analysis depending on the accuracy result and the pros and cons of each algorithm, as shown in Tables 1 and 2.

Table 1. A comparative study of ML algorithms used for blood diseases diagnosis of the last 4 years of published papers

Reference	Research Scope	Algorithms	F1 Measure	AUC	Precision
Alsheref., 2019 [8]	CBC Data	Naïve Bayes	0.835	81.60	0.862
		Bayesian Network	0.93	92.86	0.936
		Multilayer Perceptron	0.918	91.80	0.918
		Logitboost	0.98	98.16	0.982
		Random Forests	0.969	97.12	0.971
		Support Vector Machine	0.64	71.20	0.799
		K-Nearest Neighbor	0.927	92.97	0.928
		Regression Analysis	0.964	96.54	0.965
		Decision Tree	0.969	97.00	0.969
Jaiswal, 2019 [4]	CBC Data	Random Forest		95.3241	
		Naive Bayes		96.0909	
		C4.5		95.4602	
Ibrahim., 2020 [12]	Loan Prediction	Logistic Regression	0.62	0.67	0.61
		Random Forest	0.64	0.71	0.68
		Adaboost	0.63	0.72	0.66
		XGBoost	0.65	0.75	0.68
		Neural Network	0.73	0.66	0.66
		Gradient Boosting	0.66	0.75	0.68
		CatBoost	0.75	0.78	0.83
		Decision Tree	0.054	0.62	0.66

(Continued)

Table 1. A comparative study of ML algorithms used for blood diseases diagnosis of the last 4 years of published papers *(Continued)*

Reference	Research Scope	Algorithms	F1 Measure	AUC	Precision
Ibrahim, 2020 [12]	Staff Promotion	Random forest	0.94	0.71	0.70
		XGBoost	0.92	0.82	0.93
		Gradient Boost	0.95	0.82	0.93
		CatBoost	0.95	0.82	0.91
Uddin, 2022 [13]	Construction Projects Costs	Support Vector Machine	65.00	76.47	70.27
		Logistic Regression	60.00	70.59	64.86
		K-Nearest Neighbors	65.00	76.47	70.27
		Random Forest	68.18	88.24	76.92
		Stacking (Ensemble) Model	63.16	70.59	66.67
		Artificial Neural Network	65.00	76.47	70.27
Lien, 2022 [14]	CBC/DC Data	Random Forest		0.80.2	
		Logistic Regression		0.772	
Vohra., 2022 [10]	CBC Data	Decision Tree		84.72	
		Logistic Regression		94.44	
		Multilayer Perceptron		94.44	
		Naïve Bayes		87.5	
		Random Forest		88.88	
		Support Vector Machine		88.88	

To identify the best algorithms for our analysis, we conducted a comprehensive review of 6 research studies from various fields published within the last four years. Based on this analysis, we selected the best algorithms for each group, presented in Table 2. By drawing from a diverse range of research studies, we were able to identify the most effective algorithms for our purposes and ensure that our analysis was based on the latest and most reliable research findings.

Table 2. Pros and cons for the best algorithms depending on Table 1

Algorithm	Advantages	Disadvantages
Logitboost	LogitBoost supports both binary and multi-class classification. A slightly modified version of AdaBoost to handle noisy data. Applied effectively in several fields, including medical science. It classifies unidentified records pretty quickly [15] [16].	Not applied for several real-word Problems because of its complexity. Determining the linear relationships between independent and dependent variables is the main drawback [15] [16].
Random Forests	Viral algorithm used with tabular data Provide High accuracy. Very stable (work effectively with dynamic data). There is no need for normalization or standardization Less effected by noise [5] [13] [17] [18].	Need a preprocessing step (filling missing values, converting categorical to numerical data, data mapping to the same range) Complexity because of the number of trees and the method used to making the final decision [13] [17].

(Continued)

Table 2. Pros and cons for the best algorithms depending on Table 1 *(Continued)*

Algorithm	Advantages	Disadvantages
Decision Tree	Less complexity than the Random forest algorithm. Very effective when used for feature selections. Easy to understand [19] [20] [21].	Unstable. Limited Regression Performance. Overfitting problem. Effected by outliers. Less effective with extensive data. Less training time compared with the random forest but still take a lot of time for training operation [19] [20] [21].
Naive-Bayes	Relatively fast algorithm (training and testing accomplished in one pass). Appropriate for addressing tasks involving multi-class prediction. Requires less training data. Robust to noise [21] [22].	Numerical input variables are better suited for categorical input variables. Assumes variables are independent (not suitable for real-world applications). Zero probability problem. Dependent on how to input data was preprocessed [21] [22].
XGBoost	Flexibility. Highly scalable. Don't require normalized features. Perform well with nonlinear, nonmonotonic, or clustered data [23] [24].	Overfitting problem. Not effective with sparse and unstructured data [23] [24].
Gradient Boosting	It's a boosting-like algorithm for regression. Can use to improve the convergence speed and mean absolute error [23] [25].	Over-fitting problem Very sensitive to outliers [23] [25]
CatBoost	Extremely well-known and significant recently in a variety of fields. Less need for data preprocessing. Overcome the over-fitting problem. Provide better prediction results [11].	Not applied for big data to test its efficiency [11].
Support Vector Machine	Commonly used to address various issues in real-world applications (high dimensional spaces). Fairly memory efficient [13] [15] [26].	Clear margin (hyperplane) should be provided for working effectively. Not suitable for large data sets. Effected by noise [13] [15].
K-Nearest Neighbors	Time efficient in the training phase. Easy Implementation and understanding. Scalability when adding new data. Robust to noisy training data [22] [27].	Not efficient with large datasets. Not efficient with high dimensionality. Sensitive to missing data. Need normalization and standardization. Slowly implementation because of lazy learners [22] [27].
Logistic Regression	Easy Implementation and understanding. Excellent efficiency on low-dimensional data. Flexible enough to handle either continuous or discrete information [15].	Over-fitting problem Nonlinearity problems cannot be. It does poorly when the compared attributes are correlated [15] [28].
Multilayer Perceptron	Applied to complex nonlinear problems Fairly efficient with large data. Fast predication High accuracy [28].	Complex and time-consuming to implement. Highly affected by the data quality [28]

Additionally, the most recent enhanced ensemble (boosting) methods such as LIU-Boost [29], TLUSBoost [30], SecureBoost [31], SecureBoost+ [32], MPSUBoost [33]. All the 12 algorithms described in Tables 1 and 2 will be tested to find the best Algorithm to be used for the anemia classification purposes.

2.3 Evaluation step

In both phases (training and testing), we split the Dataset using 0.25% of these data for testing and the rest for training [34]. In this paper, different types of accuracy metrics were used to evaluate each method in each phase. The standard deviation and F-score with accuracy were used in the training phase. While in the testing phase, we used (Accuracy, Recall, F1-Score, Specificity (*TNR*), Precision, Sensitivity (TPR), NValue (*NPV*), PValue (*PPV*)), as shown in Equation 2 to Equation 9 [11] [35–38]. All these metrics depended on the values of *TP*, *TN*, *FP*, *FN* in the Multi-class confusion matrix.

$$F1-Score=\frac{TP}{TP+0.5(FP+FN)} \tag{1}$$

$$Accuracy=\frac{TP(TP+TF)}{Total} \tag{2}$$

$$Total=TP+Tn+FP+FN \tag{3}$$

$$Precision=\frac{TP}{(TP+FP)} \tag{4}$$

$$N\,value\,(NPV)=\frac{TN}{(TN+FN)} \tag{5}$$

$$P\,value\,(PPV)=\frac{TP}{(TP+FP)} \tag{6}$$

$$Specificity\,(TNR)=\frac{TN}{(TN+FP)} \tag{7}$$

$$Recall=\frac{TP}{(TP+FN)} \tag{8}$$

$$Sensitivity\,(TPR)=\frac{TP}{(TP+FN)} \tag{9}$$

2.4 Dataset collection and labelling step

The data collection process was carried out under two separate categories. First, ordinary CBC test samples and pathological reports were collected from Al-Zahira Hospital, comprising unclassified CBC information for 500 patients, with each record composed of 21 different features [39]. Second, for Hematology disease, CBC tests were collected from a Hematology Center in the Medical City, comprising unclassified CBC information for 300 patients, with each record composed of 28 different features [40]. The datasets were collected from these hospitals and have been made publicly available on the Mendeley datasets website, complete with a unique DOI, for wider use.

Following the completion of the data collection process, we applied the XGBoot algorithm to classify the data. This algorithm was selected based on its superior performance in analyzing the two international datasets we examined, which confirmed its accuracy and effectiveness in our research. By utilizing the best algorithm for this task, we were able to draw meaningful insights from the data.

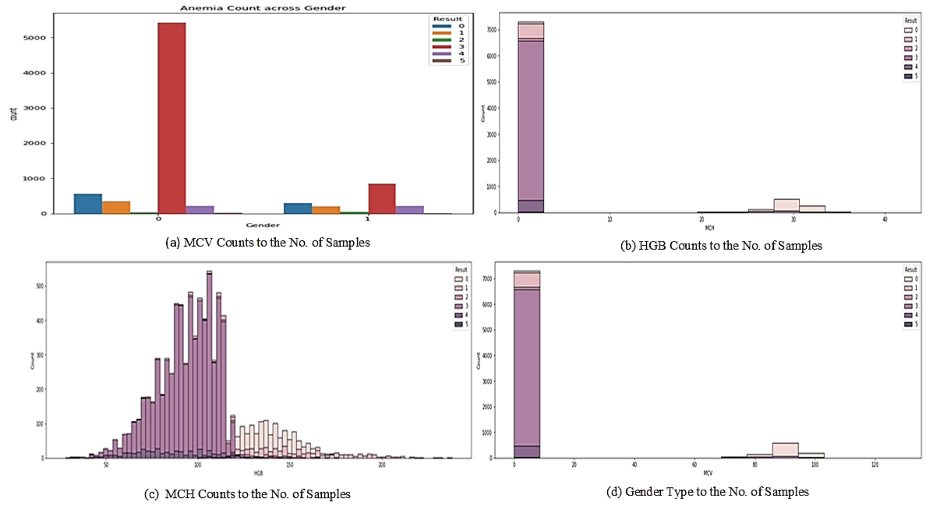
3 Results and discussions

This paper used two international datasets to learn and test the traditional methods. First, we selected the best algorithm to classify the newly built datasets to ensure it will maintain the accuracy of the data results. These methods were implemented using python3, and the datasets used will be illustrated in Table 3.

Table 3. The international tested datasets descriptions

No.	Author	Dataset Names	Year	Country	No. of Samples	No. of Features
1	Shahane S. [41]	Dataset1	2021	India	1442	5
2	Luchinin A. S.	Dataset2	2022	Russia	8246	28

The analysis process generated a wealth of information, including a range of images and data that can be used to provide several information to the medical professionals and programmers. One such example is Figure 2, which presents various statistical insights about dataset 2, the largest multi-class dataset in our study. This figure provides important information on the distribution of the dataset and the range of values for different features, which can be valuable for identifying patterns and developing effective treatment plans. The images and data generated through our analysis offer a powerful tool for medical professionals and programmers to gain deeper insights into the data, improve their understanding of medical conditions, and ultimately provide better care to patients.



(a) MCV Counts to the No. of Samples

(b) HGB Counts to the No. of Samples

(c) MCH Counts to the No. of Samples

(d) Gender Type to the No. of Samples

Fig. 2. Several examples of the statistical information of dataset2

Also, the results of the training phase using the traditional methods applied to dataset1, which contains two classes of anemia or not and has 1442 records, can be described in Figure 3. On the other hand, the results of using the same methods during the testing phase as described in Table 4.

When applying the traditional algorithms on dataset1. We can notice that the (Logitboost, Random Forest (RF), Decision Tree (DT), XGBoost, Gradient Boosting, Multilayer Perceptron, AdaBoost, Logistic Regression) will achieve the best accuracy in both training and testing phases (100%) as shown in Table 4. The rationale behind these outcomes is the efficiency of these methods with data that have a few numbers of features.

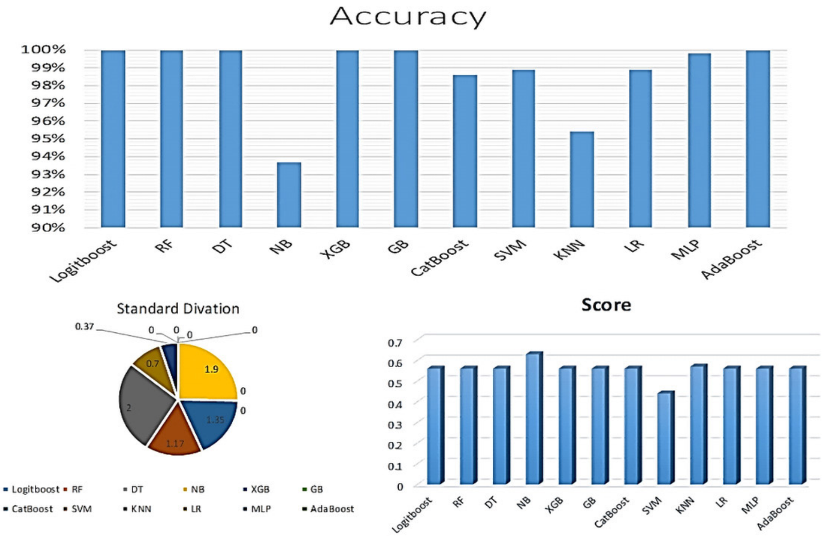


Fig. 3. Accuracy metrics for training phase using dataset1 using 10-k fold cross validation

Table 4. Accuracy metrics for 25% form the original dataset1 for testing

Algorithm	Accuracy	Recall	Score	Specificity	Precision	Sensitivity	N Value	P Value
Logitboost	1	1	66.66	1	1	1	1	1
RF	1	1	66.66	1	1	1	1	1
DT	1	1	66.66	1	1	1	1	1
NB	93.82	96.95	65.97	89.93	92.27	96.95	95.97	92.27
XGB	1	1	66.66	1	1	1	1	1
GB	1	1	66.66	1	1	1	1	1
CatBoost	99.43	99.04	66.45	1	1	99.04	98.65	1
SVM	99.71	1	66.66	99.33	99.51	1	1	99.51
KNN	97.19	99.00	66.44	94.83	96.13	99.00	98.65	96.13
LR	99.43	1	66.66	98.67	99.03	1	1	99.03
MLP	1	1	66.66	1	1	1	1	1
AdaBoost	1	1	66.66	1	1	1	1	1

In dataset 2, which consists of 8246 records divided into five different classes, we applied traditional algorithms to achieve the best accuracy during the testing phase. Our analysis revealed that the (Logitboost, Random Forest, LR, XGBoost, and Multilayer Perceptron) algorithms performed exceptionally well, as indicated by the results in Table 5. During the training phase, accuracy metrics were assessed and the best algorithms were found to be (Logitboost, Random Forest, XGBoost, Multilayer Perceptron, Logistic Regression), which achieved an accuracy rate of between 97% and 98%, as illustrated in Figure 4. However, the cat-boost algorithm was excluded from our analysis due to its inability to effectively handle large data sizes, which is considered one of its drawbacks. By utilizing the top-performing algorithms, we can more accurately and effectively analyze dataset 2, enabling us to draw more reliable conclusions about its underlying trends and patterns.

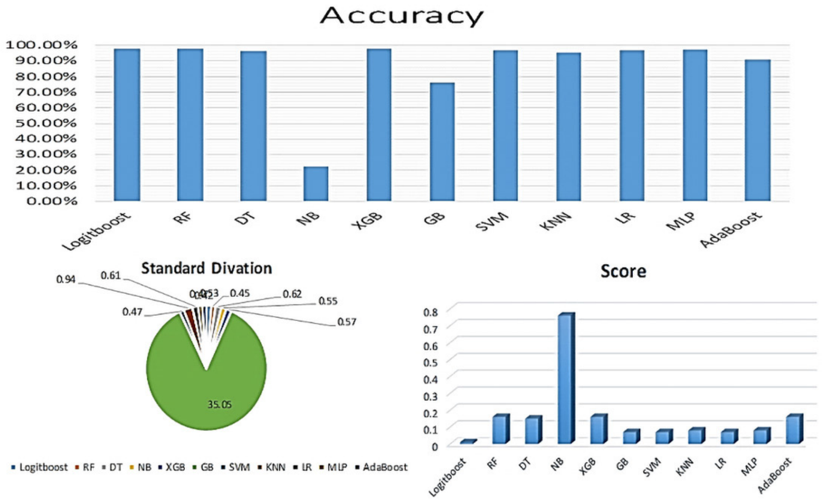


Fig. 4. Accuracy metrics for training phase using dataset 2 using 10-k fold cross validation

Table 5. Accuracy metrics for 25% form the original dataset2 for testing

Algorithm	Accuracy	Recall	Score	Specificity	Precision	Sensitivity	N Value	P Value
Logitboost	97.81	89.96	64.27	99.55	97.81	89.96	97.81	97.81
RF	97.13	87.16	63.54	99.41	97.13	87.16	97.13	97.13
DT	96.36	84.12	62.72	99.25	96.36	84.12	96.36	96.36
NB	20.99	05.04	09.16	57.06	20.99	05.04	20.99	20.99
XGB	98.25	91.84	64.74	99.64	98.25	91.84	98.25	98.25
GB	95.05	79.35	61.34	98.96	95.05	79.35	95.05	95.05
SVM	96.84	86.00	63.23	99.35	96.84	86.00	96.84	96.84
KNN	95.44	80.72	61.75	99.05	95.44	80.72	95.44	95.44
LR	97.09	86.96	63.49	99.40	97.09	86.96	97.09	97.09
MLP	97.67	89.35	64.11	99.52	97.67	89.35	97.67	97.67
AdaBoost	91.07	67.11	57.30	98.07	91.07	67.11	91.07	91.07

In order to ensure unbiased testing results, we have increased the size of our testing set to 30%. Based on literature [42] and [43], this percentage is optimal for obtaining the best model performance. When applying this percentage to dataset1, we found that the results were almost identical to those achieved with a 25% testing set size. However, when applying the 30% size to dataset2, the results showed a noticeable difference, as illustrated in Figure 5 and Table 6. By increasing the testing set size, we can more effectively evaluate the performance of our models and obtain more reliable results, which will ultimately improve the accuracy of our findings.

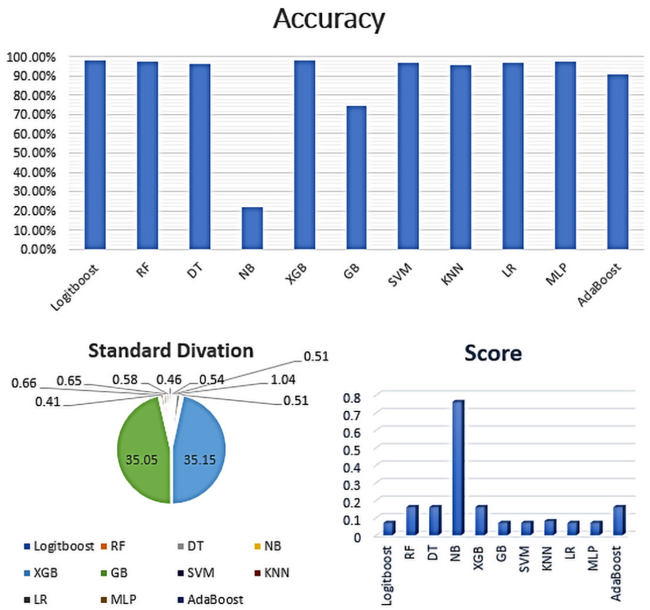


Fig. 5. Accuracy metrics for training phase using dataset2 using 10-k fold cross validation

Table 6. Accuracy metrics for 30% form the original dataset2 for testing

Algorithm	Accuracy	Recall	Score	Specificity	Precision	Sensitivity	N Value	P Value
Logitboost	97.57	97.57	66.11	97.57	99.50	97.57	88.94	99.50
RF	97.49	97.49	66.10	97.49	99.48	97.49	88.61	99.48
DT	96.25	96.11	65.78	96.94	99.37	96.11	83.20	99.37
NB	20.97	20.97	29.55	20.97	57.032	20.97	5.04	57.03
XGB	98.05	98.05	66.22	98.05	99.60	98.05	90.99	99.60
GB	90.20	95.71	65.68	70.03	92.11	95.71	81.71	92.11
SVM	96.92	96.92	65.96	96.92	99.37	96.92	86.32	99.37
KNN	95.63	95.63	65.66	95.63	99.09	95.63	81.41	99.09
LR	96.92	96.92	65.96	96.92	99.37	96.92	86.32	99.37
MLP	97.45	97.45	66.09	97.45	99.48	97.45	88.44	99.48
AdaBoost	90.86	90.86	64.50	90.86	98.02	90.86	66.54	98.02

After conducting a comparative study of the two datasets of varying sizes, we observed that the best performance was achieved by (Logitboost, Random Forest, SVM, XGBoost, Multi-layer Perceptron) algorithms. To leverage the strengths of these algorithms and minimize their weaknesses, we can create a new hybrid model. In this paper, we selected XGBoost as the best-performing algorithm across both datasets. While Random Forest achieved the best results in datasets 1 and 2, its performance was still good in dataset 2. In cases where the SVM and MLP are closely performing, we should choose MLP, as the SVM algorithm is sensitive to outlier data and it is not suitable for large data sets. By selecting the most appropriate algorithms and combining them into a new hybrid model, we can more effectively analyze datasets and obtain more reliable and accurate results.

As a final step for this research, the XGBoost classifier was used to classify the anemia disease of newly collected dataset into multiple anemia types (Normal CBC, Myeloproliferative Disorder, Chronic Lymphocytic Leukemia, Iron Deficiency Anemia, clinically significant, and Not Clinically Significant). We apply the XGBoost Classifier to the Hematology dataset (300 Sample) after training the classifier with dataset2 (Dataset from Russia); the results are shown in Figure 6.

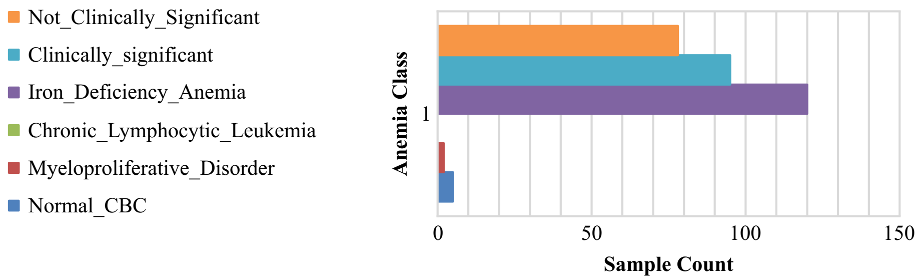


Fig. 6. The anemia classes for the hematology dataset

The results show that the most significant number of the samples suffer from Iron Deficiency Anemia (IDA). The IDA means the patient suffers from anemia due to iron deficiency. On the other hand, Figure 7 shows the results when we apply the XGBoost to the CBC dataset.

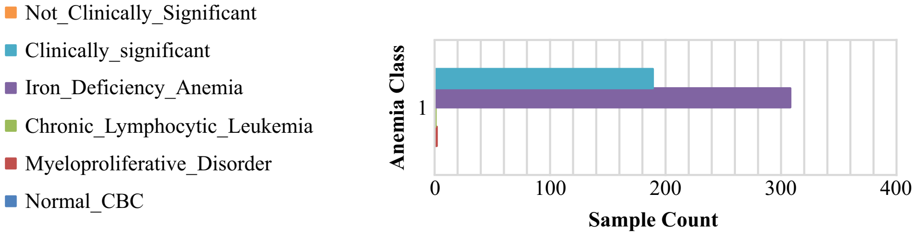


Fig. 7. The anemia classes for the CBC dataset

The classification results also show that the most significant number of the samples classified as Iron Deficiency Anemia (IDA). While, the second large number of patients were classified as a Clinically Significant, this group divided into IDA, Acute Blood Diseases (acute leukemia, Myelodysplastic syndromes, Aplastic Anemia), myeloproliferative disorder, and Chronic lymphocytic leukaemia.

4 Conclusions

In conclusion, the use of machine learning (ML) models for anemia disease classification has shown promising results, as demonstrated by the testing of 12 different ML algorithms. That's where all of the Logitboost, Random Forest, XGBoost, and Multilayer Perceptron achieved an acceptable value of accuracy in experiments conducted on two international datasets. As result the XGBoost classifier provides the best accuracy for all tested datasets 100% and 97.81% for dataset1, and dataset2, respectively. This suggests that ML models can be effective tools for identifying and classifying medical conditions in anemia, which could improve diagnostic accuracy and potentially lead to better treatment outcomes.

However, it is important to note that the effectiveness of ML models is highly dependent on the quality and quantity of the data used to train them. Therefore, it is crucial to ensure that the data used for training is diverse, representative, and of high quality. Moreover, ML models are not foolproof and may still make errors, especially when dealing with complex and nuanced medical conditions.

Despite these limitations, the potential applications of ML models in healthcare are vast, and the use of these models for anemia disease classification is an exciting development in this field. Further research and validation are necessary to evaluate the effectiveness of ML models for anemia disease classification, and to explore potential applications in other areas of healthcare. Nonetheless, these initial findings provide a strong foundation for future investigations and highlight the significant potential of ML models in improving medical diagnoses and treatments.

5 References

[1] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245):255–260. https://doi.org/10.1126/science.aaa8415

[2] Kadhim, Z. S., Abdullah, H. S., & Ghathwan, K. I. (2022). Artificial Neural Network hyperparameters optimization: A survey. International Journal of Online and Biomedical Engineering, 18(15):59–87. https://doi.org/10.3991/ijoe.v18i15.34399

[3] Hafeez Kandhro, A., Shoombuatong, W., Prachayasittikul, V., & Nuchnoi, P. (2017). New bioinformatics-based discrimination formulas for differentiation of thalassemia traits from iron deficiency anemia. Laboratory Medicine, 48:230–237. https://doi.org/10.1093/labmed/lmx029

[4] Jaiswal, M., Srivastava, A., & Siddiqui, T. J. (2019). Machine learning algorithms for anemia disease prediction. In Recent trends in communication, computing, and electronics, Springer, Singapore, pp. 463–469. https://doi.org/10.1007/978-981-13-2685-1_44

[5] Lippi, G., Mattiuzzi, C., & Cervellin, G. (2018). Artificial intelligence and clinical laboratory medicine. Clinical Chemistry and Laboratory Medicine, 56(12):2000–2007. https://doi.org/10.1515/cclm-2018-0312

[6] Li, N., Li, Y., Li, X., Guo, Y., & Li, Y. (2020). Prediction model of iron deficiency anemia in pregnant women based on machine learning. Clinical Epidemiology, 12:303–311.

[7] Yan, L., Yao, Y., Wang, P., et al. (2019). A machine learning-based model for predicting anemia in patients with chronic kidney disease. Scientific Reports, 9(1):18519.

[8] Alsheref, F. K., & Gomaa, W. H. (2019). Blood diseases detection using classical machine learning algorithms. International Journal of Advanced Computer Science and Applications, 10(7):77–81. https://doi.org/10.14569/IJACSA.2019.0100712

[9] Haider, R. Z., Ujjan, I. U., Khan, N. A., Urrechaga, E., & Shamsi, T. S. (2022). Beyond the In-Practice CBC: The research CBC parameters-driven machine learning predictive modeling for early differentiation among leukemias. Diagnostics, 12(1):138. https://doi.org/10.3390/diagnostics12010138

[10] Vohra, R., Hussain, A., Dudyala, A. K., Pahareeya, J., & Khan, W. (2022). Multi-class classification algorithms for the diagnosis of anemia in an outpatient clinical setting. Plos One, 17(7):e0269685. https://doi.org/10.1371/journal.pone.0269685

[11] Abdul-Jabbar, S. S., Farhan, A. K., Abdelhamid, A. A., & Ghoneim, M. E. (2022). Razy: A string matching algorithm for automatic analysis of pathological reports Axioms, 11(10):547. https://doi.org/10.3390/axioms11100547

[12] Ibrahim, A. A., Ridwan, R. L., Muhammed, M. M., Abdulaziz, R. O., & Saheed, G. A. (2020). Comparison of the CatBoost classifier with other machine learning methods. International Journal of Advanced Computer Science and Applications, 11(11):738–748. https://doi.org/10.14569/IJACSA.2020.0111190

[13] Uddin, S., Ong, S., & Lu, H. (2022). Machine learning in project analytics: A data-driven framework and case study. Scientific Reports, 12(1):1–13. https://doi.org/10.1038/s41598-022-19728-x

[14] Lien, F., Lin, H. S., Wu, Y. T., & Chiueh, T. S. (2022). Bacteremia detection from complete blood count and differential leukocyte count with machine learning: Complementary and competitive with C-reactive protein and procalcitonin tests. BMC infectious diseases, 22(1):1–10. https://doi.org/10.1186/s12879-022-07223-7

[15] Pham, B. T., & Prakash, I. (2019). Evaluation and comparison of LogitBoost ensemble, fisher’s linear discriminant analysis, logistic regression and support vector machines methods for landslide susceptibility mapping. Geocarto International, 34(3):316–333. https://doi.org/10.1080/10106049.2017.1404141

[16] Cai, Y. D., Feng, K. Y., Lu, W. C., & Chou, K. C. (2006). Using LogitBoost classifier to predict protein structural classes. Journal of Theoretical Biology, 238(1):172–176. https://doi.org/10.1016/j.jtbi.2005.05.034

[17] Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. The Stata Journal, 20(1):3–29. https://doi.org/10.1177/1536867X20909688

[18] Hussein, A. Y., & Sadiq, A. T. (2022). Meerkat clan-based feature selection in random forest algorithm for IoT intrusion detection. Iraqi Journal of Computers, Communications, Control and Systems Engineering, 22(3):15–24.

[19] M. Hui, W. W., Zhang, B., Scherer, R., & Damaševičius, R. (2021). Research on decision tree based on rough set. Journal of Internet Technology, 22(6):1385–1394. https://doi.org/10.53106/160792642021112206015

[20] Mittal, K., Khanduja, D., & Tewari, P. C. (2017). An insight into 'Decision Tree Analysis'. World Wide Journal of Multidisciplinary Research and Development, 3(12):111–115.

[21] Kalcheva, N., Todorova, M., & Marinova, G. (2020). Naive bayes classifier, decision tree and AdaBoost ensemble algorithm–advantages and disadvantages. Knowledge Based Sustainable Development, (2020):153. https://doi.org/10.31410/ERAZ.2020.153

[22] Jadhav, S. D., & Channe, H. P. (2016). Comparative study of K-NN, naive bayes and decision tree classification techniques. International Journal of Science and Research (IJSR), 5(1):1842–1845. https://doi.org/10.21275/v5i1.NOV153131

[23] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2019). A Comparative Analysis of XGBoost., arXiv:1911.01914.

[24] Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: An experimental review. Journal of Big Data, 7(1):1–47. https://doi.org/10.1186/s40537-020-00349-y

[25] Ali, A. T., Abdullah, H. S., & Fadhil, M. N. (2021). Impostor recognition based voice authentication by applying three machine learning algorithms. Iraqi Journal of Computers, Communications, Control and Systems Engineering, 21(3):112–124. https://doi.org/10.33103/uot.ijccce.21.3.10

[26] Mustapha, A., Mohamed, L., & Ali, K. (2021). Comparative study of optimization techniques in deep learning: Application in the ophthalmology field. In Journal of Physics: Conference Series, IOP Publishing, 1743(1):012002. https://doi.org/10.1088/1742-6596/1743/1/012002

[27] Puspitasari, N., Burhandeny, A. E., Nurulita, A. D. A., & Trahutomo, D. (2022). Naïve bayes and K-nearest neighbor algorithms performance comparison in diabetes mellitus early diagnosis. International Journal of Online and Biomedical Engineering, 18(15):202–215. https://doi.org/10.3991/ijoe.v18i15.34143

[28] Çolakoğlu, N., & Akkaya, B. (2019). Recent Advances in Data Science and Business Analytics. Mimar Sinan Fine Arts University Publications, y-BIS 2019 Conference Book: 884.

[29] Ahmed, S., Rayhan, F., Mahbub, A., Jani, R., Shatabda, S., & Farid, D. M. (2019). LIUBoost: locality informed under-boosting for imbalanced data classification. In Emerging Technologies in Data Mining and Information Security: pp. 133–144, Springer, Singapore. https://doi.org/10.1007/978-981-13-1498-8_12

[30] Kumar, S., Biswas, S. K., & Devi, D. (2019). TLUSBoost algorithm: A boosting solution for class imbalance problem. Soft Computing, 23(21):10755–10767. https://doi.org/10.1007/s00500-018-3629-4

[31] Cheng, K., Fan, T., Jin, Y., Liu, Y., Chen, T., Papadopoulos, D., & Yang, Q. (2021). Secureboost: A lossless federated learning framework. IEEE Intelligent Systems, 36(6):87–98. https://doi.org/10.1109/MIS.2021.3082561

[32] Chen, W., Ma, G., Fan, T., Kang, Y., Xu, Q., & Yang, Q. (2021). SecureBoost+: A high performance gradient boosting tree framework for large scale vertical federated learning. arXiv preprint arXiv:2110.10927.

[33] Kim, S. J., & Lim, D. J. (2022). MPSUBoost: A modified particle stacking undersampling boosting method. IEEE Access, 10:125458–125468. https://doi.org/10.1109/ACCESS.2022.3225456

[34] El-kenawy, E.S.M.T., 2019. A machine learning model for hemoglobin estimation and anemia classification. International Journal of Computer Science and Information Security (IJCSIS), 17(2):100–108.

[35] Kılınç, D. (2016). An accurate toponym-matching measure based on approximate string matching. Journal of Information Science, 42(2):138–149. https://doi.org/10.1177/0165551515590097

[36] El-Kenawy, E. S. M., Mirjalili, S., Alassery, F., Zhang, Y. D., Eid, M. M., El-Mashad, S. Y. & Abdelhamid, A. A. (2022). Novel meta-heuristic algorithm for feature selection, unconstrained functions and engineering problems. IEEE Access, 10:40536–40555. https://doi.org/10.1109/ACCESS.2022.3166901

[37] Ali, L. R., Jebur, S. A., Jahefer, M. M., & Shaker, B. N. (2022). Employing transfer learning for diagnosing COVID-19 disease. International Journal of Online and Biomedical Engineering, 18(15):31–42. https://doi.org/10.3991/ijoe.v18i15.35761

[38] Errabih, A., Boussarhane, M., Nsiri, B., Sadiq, A., Alaoui, M. H. E. Y., Thami, R. O. H., & Benaji, B. (2022). Identifying retinal diseases on OCT image based on deep learning. International Journal of Online and Biomedical Engineering, 18(15):141–159. https://doi.org/10.3991/ijoe.v18i15.33639

[39] Abdul-Jabbar, S. S., & Farhan, A. (2022). CBC Dataset. Mendeley Data. V1., https://doi.org/10.17632/28s2bhdjfd.1

[40] Abdul-Jabbar, S. S., & Farhan, A. (2022). Hematological Dataset, Mendeley Data, V1. https://doi.org/10.17632/g7kf8x38ym.1

[41] Shahane, S. (2020). Anemia Diagnosis Dataset. India. [Online]. Available: https://www.kaggle.com/datasets/biswaranjanrao/anemia-dataset?select=anemia.csv

[42] Nguyen, Q. H., Ly, H. B., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., … & Pham, B. T. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. Mathematical Problems in Engineering, 2021:1–15. https://doi.org/10.1155/2021/4832864

[43] Mohammed, Z. A., Abdullah, M. N., & Al Hussaini, I. H. (2021). Predicting incident duration based on machine learning methods. Iraqi Journal of Computers, Communications, Control and Systems Engineering, 21(1):1–15. https://doi.org/10.33103/uot.ijccce.21.1.1

6 Authors

Safa S. Abdul-Jabbar is an Assistant Lecturer at the Department of Computer Science, College of Science for Women, University of Baghdad, Baghdad, Iraq. (email: safa.s@csw.uobaghdad.edu.iq).

Dr. Alaa K. Farhan a Professor at the Department of Computer Sciences, University of Technology, Baghdad, Iraq. He got his PhD. degree in Information Security from the University of Technology, Baghdad, in 2009. (email: alaa.k.farhan@uotechnology.edu.iq).

Dr. Alexander S. Luchinin (MD, PhD), Federal State Budgetary Institution of Science "Kirov Research Institute of Hematology and Blood Transfusion of the Federal Medical and Biological Agency", Russia, 610027, Kirov region, Kirov, Krasnoarmeyskaya, 72. (email: glivec@mail.ru).

Article submitted 2023-01-18. Resubmitted 2023-02-24. Final acceptance 2023-02-24. Final version published as submitted by the authors.