Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 8, S. 108-119, 7, 2024

Araştırma Makalesi

IJANSER

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 8, pp. 108-119, 7, 2024

Research Article

https://as-proceeding.com/index.php/ijanser
ISSN: 2980-0811

Evaluating Optimizable Machine Learning Models for Anemia Type Prediction from Complete Blood Count Data

Ladislav Végh [1*], Ondrej Takáč [1], Krisztina Czakóová [1], Daniel Dancsa [2] and Melinda Nagy [2]

[1] *Department of Informatics, Faculty of Economics and Informatics, J. Selye University, Slovakia*
[2] *Department of Biology, Faculty of Education, J. Selye University, Slovakia*

[*]*(veghl@ujs.sk) Email of the corresponding author*

(Received: 21 August 2024, Accepted: 28 August 2024)

(5th International Conference on Engineering and Applied Natural Sciences ICEANS 2024, August 25-26, 2024)

ATIF/REFERENCE: Végh, L., Takáč, O., Czakóová, K., Dansca, D. & Nagy, M. (2024). Evaluating Optimizable Machine Learning Models for Anemia Type Prediction from Complete Blood Count Data. *International Journal of Advanced Natural Sciences and Engineering Researches*, 8(7), 108-119.

Abstract – This paper compares different optimizable machine learning classification models to predict eight types of anemia from complete blood count (CBC) data. For the research, we used a publicly available Kaggle dataset containing 1281 observations, 14 predictors, and the diagnosis as the categorical target variable with nine categories (eight types of anemia and the healthy category). First, we examined the dataset and observed the histograms of some of the predictors. We compared the values of predictors of observations with no anemia to the observations where any anemia was diagnosed. Next, we used MATLAB R2024a to train and test nine optimizable machine-learning classification models. These models were Ensemble, Tree, SVM, Efficient Linear, Neural Network, Kernel, KNN, Naïve Bayes, and the Discriminant. Bayesian optimization was used to optimize the hyperparameters of all these models. We used 90% of observations for training and 10% of observations for testing. During the training, 10-fold cross-validation was used to prevent overfitting. The results showed the best accuracy was reached with the Ensemble classification model using the bag ensemble method (validation accuracy: 99.22%, test accuracy: 100%). Finally, we inspected our best classification model in more detail. We calculated the permutation feature importance to determine the contribution of each predictor to the final model. The results showed 6–7 important predictors, while the most important feature was the amount of hemoglobin.

Keywords – Data Exploration, Machine Learning, Multiclass Classification, Anemia Types, Complete Blood Count, CBC Test.

I. INTRODUCTION

Artificial intelligence, deep learning, and machine learning can be used in various fields to recognize patterns from data [1], [2], [3], [4]. Many deep learning and machine learning classification models can be efficiently used in healthcare, as well, e.g., for diagnosis of different types of cancer [5], predicting diabetes [6], [7], cardiovascular disease [8], [9], or anemia [10], [11], [12]. In this paper, we use machine learning techniques to diagnose various anemia types from complete blood count (CBC) data.

Many of the surveys related to diet and nutrition uncovered that almost a quarter of the world's population is anemic [12]. Typical symptoms of anemia include fatigue, weakness, and shortness of breath, caused mainly by low hemoglobin levels or insufficient oxygen-carrying capacity. Timely and correct diagnosis, followed by appropriate treatment, are vital steps in curing anemia [10], [13].

Traditionally, hematologists manually examine blood tests to diagnose different types of anemia. However, this process might be time-consuming and prone to human errors, which might delay the proper treatment [10], [11]. Furthermore, discriminating iron deficiency anemia (IDA) from other types of anemia requires a more expensive test (serum ferritin) than the CBC test [14]. Machine learning can help to less costly, time-savingly, and accurately diagnose various types of anemia from CBC data [14], [15].

In the following parts of this paper, we focused on exploring a dataset containing CBC data of 1281 patients. We compared the accuracy of nine optimizable machine-learning classification models for classifying eight different types of anemia. These anemia types are iron deficiency anemia, leukemia, leukemia with thrombocytopenia, macrocytic anemia, normocytic hypochromic anemia, normocytic normochromic anemia, other microcytic anemia, and thrombocytopenia. We trained, tested, and optimized various classification models to find the best model for the given dataset. Finally, we calculated permutation feature importance to determine which of the 14 predictors of the dataset are the most important.

II. MATERIALS AND METHOD

For this research, we used Anemia Types Classification [16] dataset downloaded from Kaggle. For data exploration, SPSS [17] and MATLAB [18] software were used. For training, testing, optimizing the classification models, and calculating permutation feature importance, MATLAB [18] software was used.

A. Dataset

The dataset [16] contains 1281 observations, 14 predictors (CBC data), and the diagnosis as a categorical target variable.

The 14 predictors are the amount of hemoglobin (HGB), the number of platelets (PLT), the count of white blood cells (WBC), the count of red blood cells (RBC), the hematocrit test (HCT), the mean corpuscular volume (MCV), the mean corpuscular hemoglobin (MCH), the mean corpuscular hemoglobin concentration (MCHC), the variability in platelet size distribution in the blood (PDW), the procalcitonin test (PCT), the percent of lymphocytes (LYMp), the percent of neutrophils (NEUTp), the number of lymphocytes (LYMn), and the number of neutrophils (NEUTn).

The target variable contains nine categories: one for healthy patients and eight for different types of anemia. The dataset includes the following categories for various anemia types: iron deficiency anemia, leukemia, leukemia with thrombocytopenia, macrocytic anemia, normocytic hypochromic anemia, normocytic normochromic anemia, other microcytic anemia, and thrombocytopenia.

B. Data Exploration

First, we wanted to know how the data is distributed by diagnosis. Fig. 1 shows the distribution of observations by target variable. As we can see, the dataset contains more observations for some categories and less for others. There are 336 observations in the Healthy category, 279 observations in the Normocytic hypochromic anemia category, and 269 observations in the Normocytic normochromic anemia category, but only 18 observations in the Macrocytic anemia category and 11 observations in the Leukemia with thrombocytopenia category.

Next, we were curious if there is a significant difference in any predictors between observations without anemia (healthy patients) and observations with any anemia. For this reason, we calculated the mean, standard deviation, and median values for every predictor of the two groups. Table 1 shows the results. By examining the values in this table, we can observe differences for some of the predictors, but for others, there is only a slight difference or no difference at first sight.

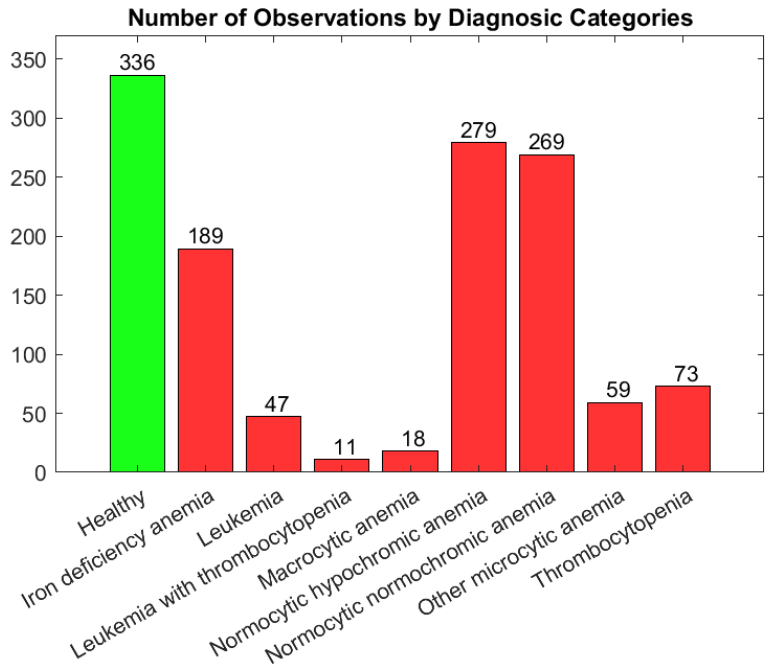

Fig. 1 Number of observations by target variable

Table 1. Comparison predictors' means, standard deviations, and medians of the observations without anemia (healthy patients) to the observations with some anemia diagnosed

Predictor	Observations without anemia (N=336)			Observations with anemia (N=945)		
	Mean	Std. Dev.	Median	Mean	Std. Dev.	Median
HGB	13.99	0.71	13.90	11.54	4.24	11.60
PLT	297.34	83.08	330.00	206.03	84.23	192.00
WBC	7.74	1.31	8.00	7.92	4.08	7.10
RBC	5.25	0.40	5.28	4.55	3.25	4.41
HCT	45.75	1.30	46.15	46.30	122.13	46.15
MCV	90.46	4.56	91.00	84.14	31.36	85.00
MCH	29.58	1.98	30.00	32.97	129.44	27.00
MCHC	32.63	3.52	32.35	31.42	3.16	31.70
PDW	14.62	2.02	14.31	14.20	3.28	14.31
PCT	0.25	0.03	0.26	0.27	0.80	0.26
LYMp	26.11	4.27	25.85	25.75	7.79	25.85
NEUTp	77.29	33.08	77.51	77.59	170.91	77.51
LYMn	1.88	0.29	1.88	1.88	1.55	1.88
NEUTn	5.04	0.66	5.14	5.18	3.32	5.14

Afterward, we used statistical tests to evaluate the significance of the differences between the two groups. We tested the normality of data using Shapiro-Wilk tests. Because the results showed no normal distribution in any of the groups for any of the predictors, we used Mann-Whitney U tests to determine if there was a significant difference in predictors between observations without anemia (group 0) and observations with any anemia (group 1). In Table 2, we can see the results of the Mann-Whitney U tests. According to the test results, group 0 (healthy patients) has significantly higher values for every predictor than group 1 (patients with anemia diagnosed), except for the predictor LYMp, where there is no significant difference between the groups.

Table 2. Results of the Mann-Whitney U tests

Predictor	Mean Rank of Group 0 (Healthy, N=336)	Mean Rank of Group 1 (Anemia, N=945)	Mann-Whitney U	Z	Asymp. Sig. (2-tailed)
HGB	1029.38	502.91	28264.000	-22.409	0.000
PLT	907.07	546.40	69362.000	-15.351	0.000
WBC	742.70	604.84	124589.500	-5.868	0.000
RBC	993.47	515.68	40329.500	-20.336	0.000
HCT	810.79	580.63	101711.500	-11.138	0.000
MCV	935.95	536.13	59657.000	-17.018	0.000
MCH	927.85	539.01	62380.000	-16.562	0.000
MCHC	841.24	569.80	91479.000	-11.618	0.000
PDW	713.32	615.29	134459.500	-4.247	0.000
PCT	777.83	592.35	112783.500	-8.978	0.000
LYMp	661.96	633.55	151717.000	-1.375	0.169
NEUTp	740.18	605.74	125437.000	-6.506	0.000
LYMn	694.45	622.00	140802.000	-3.506	0.000
NEUTn	682.43	626.27	144841.000	-2.718	0.007

Next, we wanted to see the data visually, especially the difference between the two groups; for this reason, we created a histogram for every predictor (see Fig. 2–8). In these histograms, observations of healthy patients are marked with green color, while observations of patients with any anemia are marked with red. By observing these charts, we might notice a clear visual difference between the two groups for some predictors; however, there is no noticeable difference for other predictors, even though the Man-Whitney U tests showed significant differences.

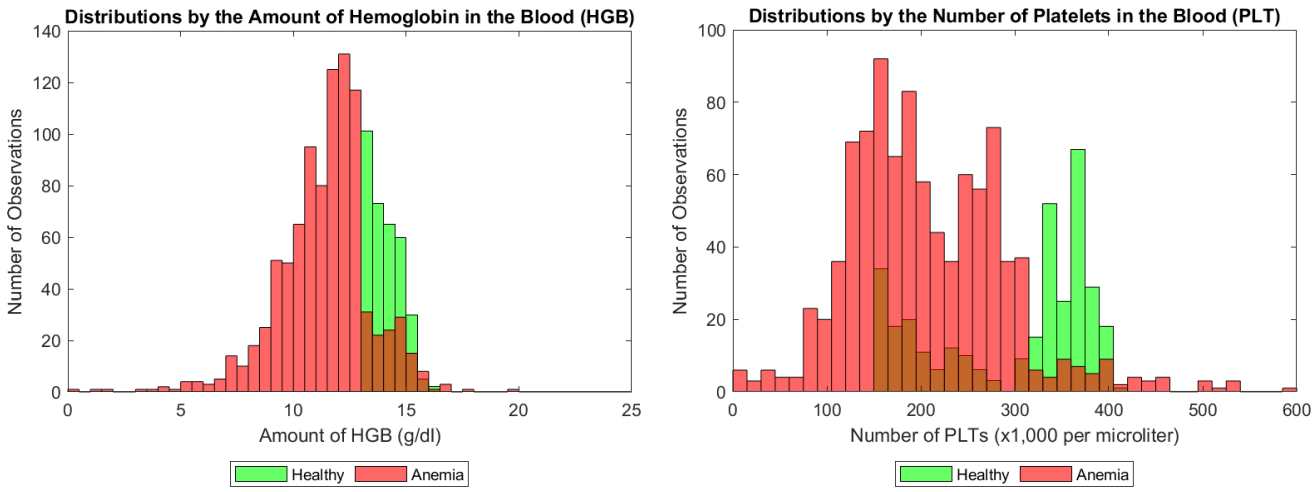

Fig. 2 Distribution by the amount of hemoglobin (HGB) and by the number of platelets in the blood (PLT)

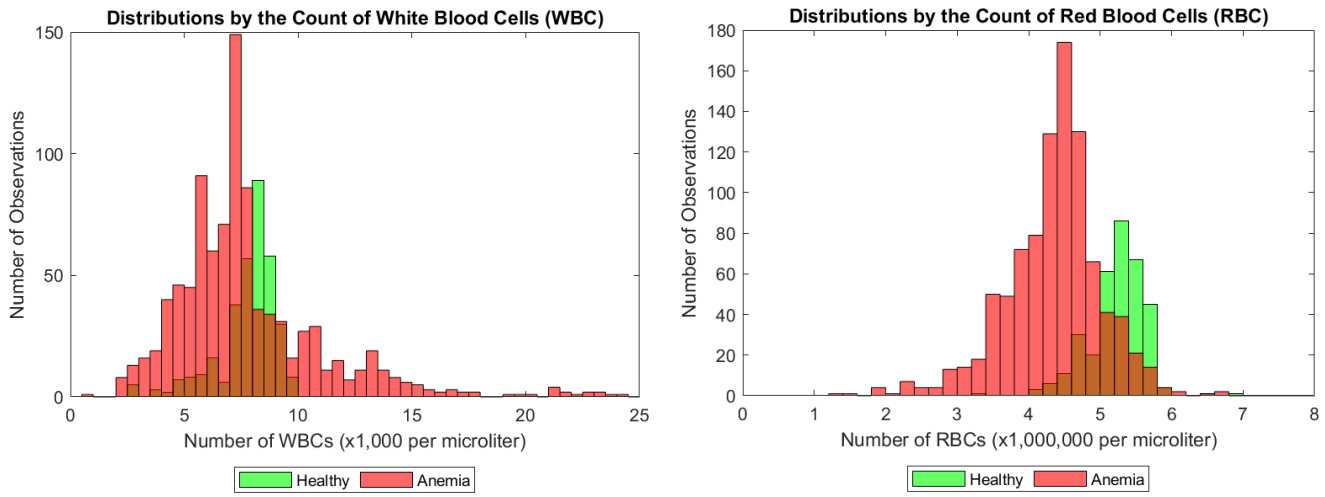

Fig. 3 Distribution by the count of white blood cells (WBC) and by the count of red blood cells (RBC)

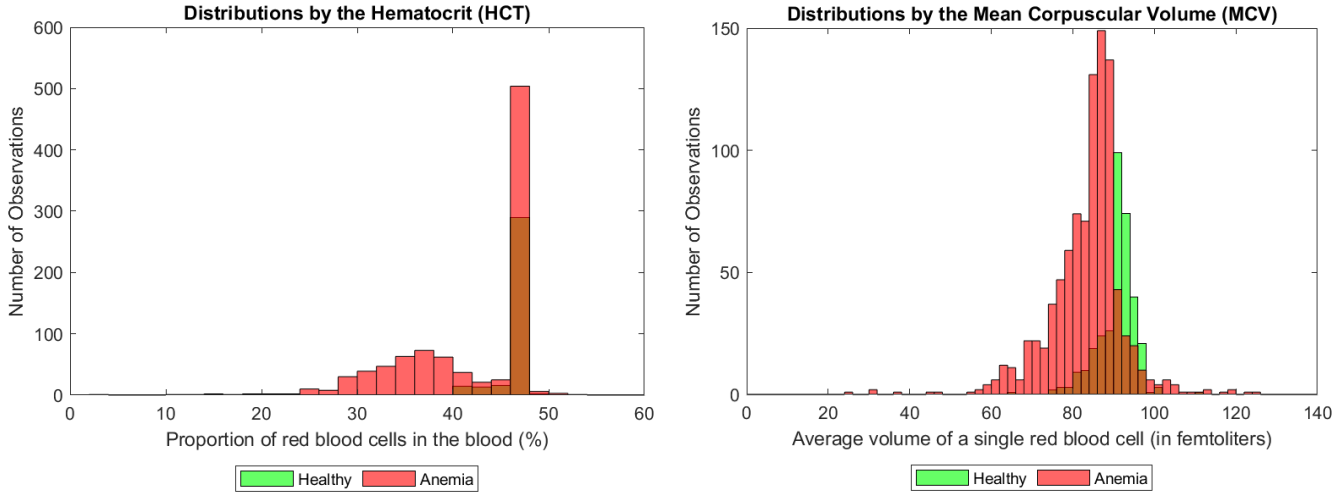

Fig. 4 Distribution by the hematocrit (HTC) and by the mean corpuscular volume (MCV)

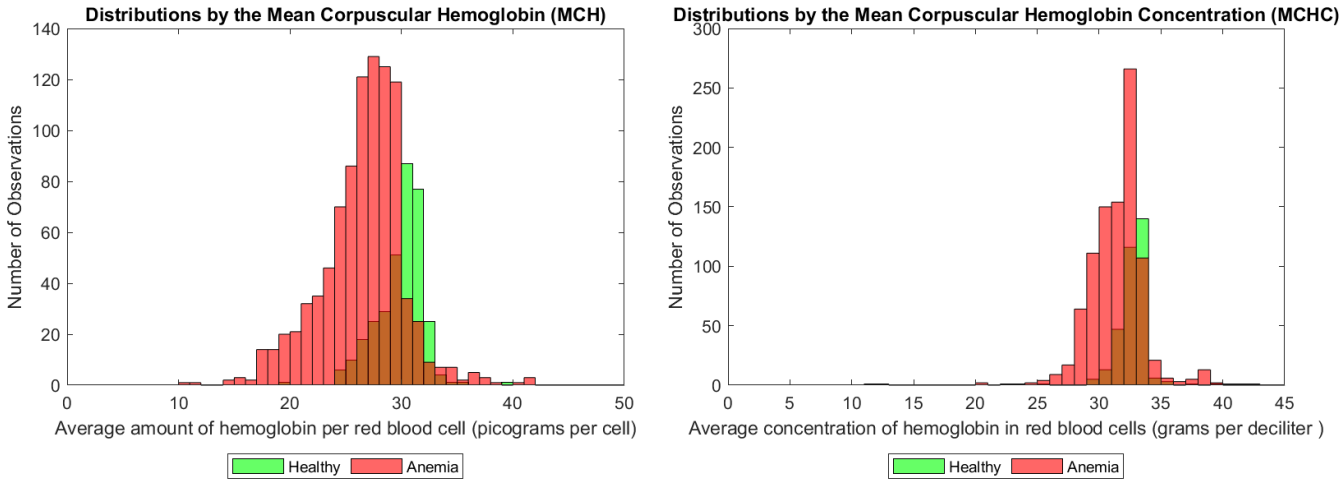

Fig. 5 Distribution by the mean corpuscular hemoglobin (MCH) and by the mean corpuscular hemoglobin concentration (MCHC)

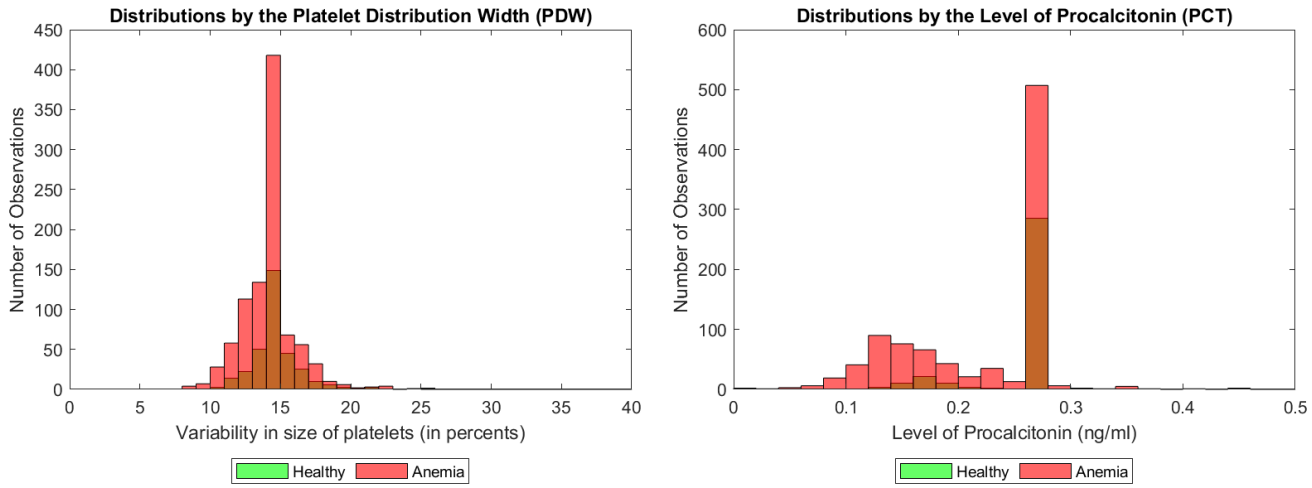

Fig. 6 Distribution by the platelet distribution width (PDW) and by the level of procalcitonin (PCT)

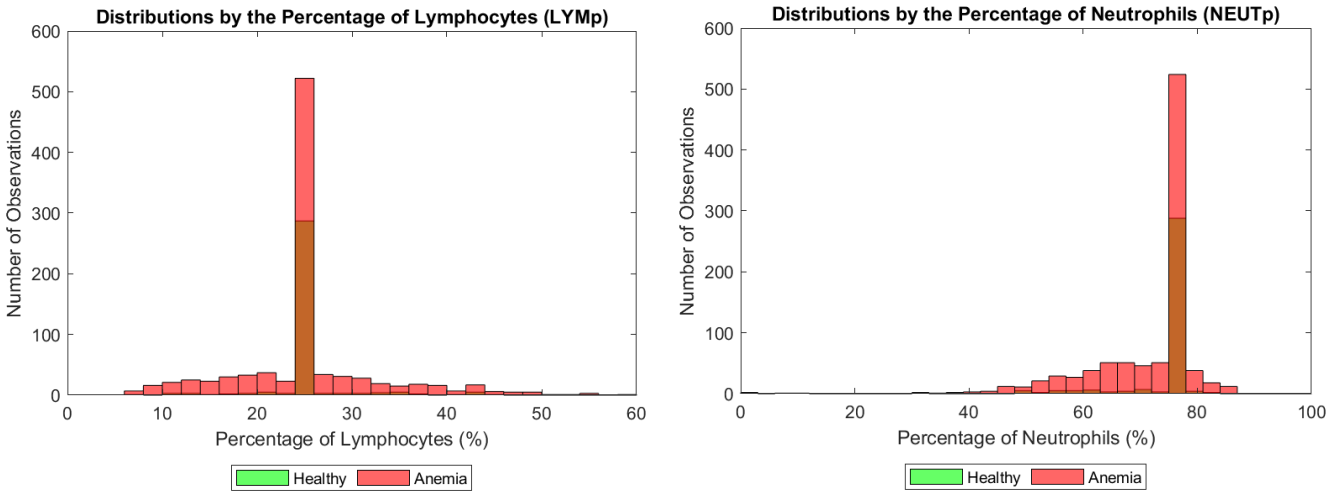

Fig. 7 Distribution by the percentage of lymphocytes (LYMp) and by the percentage of neutrophils (NEUTp)

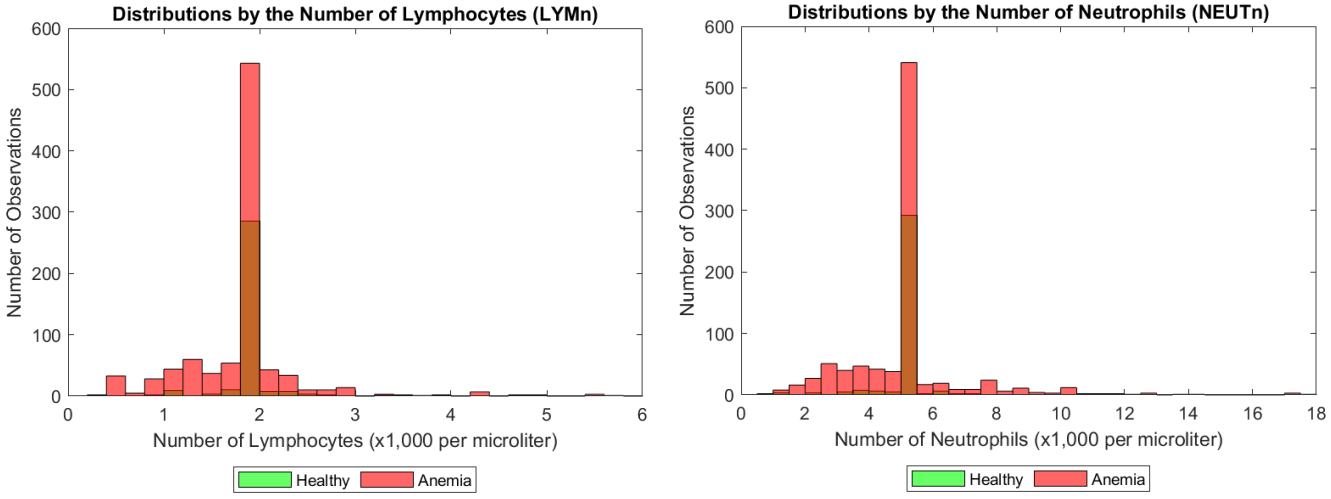

Fig. 8 Distribution by the number of lymphocytes (LYMn) and by the number of neutrophils (NEUTn)

C. *Training and Testing Classification Models*

After data exploration, our goal was to find a machine learning classification model that not only classifies observations into two groups (healthy and anemia) but also accurately classifies the eight types of anemia categories.

We split the dataset into a training set (90% of observations) and a test set (10% of observations for testing). Using the training set and MATLAB R2024a [18] we trained nine optimizable machine learning classification models in MATLAB's Classification Learner App [19]. These models were Ensemble, Tree, SVM, Efficient Linear, Neural Network, Kernel, KNN, Naïve Bayes, and Discriminant. Bayesian optimization was used to optimize the hyperparameters of all these models. During the training, 10-fold cross-validation was used to prevent overfitting. After training, the final classification models were tested on the test data set.

III. RESULTS

Table 3 shows the validation and test accuracies of the machine learning classification models. The best result was reached with the Ensemble classification model, which had a validation accuracy of 99.22% and a test accuracy of 100%.

Table 3. Validation and test accuracies of the classification models

#	Model Type	Accuracy % (Validation)	Accuracy % (Test)
1	Ensemble	99.22	100
2	Tree	99.05	100
3	SVM	91.76	91.41
4	Efficient Linear	89.51	94.53
5	Neural Network	88.03	89.06
6	Kernel	81.35	82.81
7	KNN	75.80	77.34
8	Naive Bayes	67.30	67.19
9	Discriminant	54.38	52.34

All these models were optimizable machine learning models, where Bayesian optimization was used to find the models' minimum error and best point hyperparameters. Table 4 shows the minimum error (also best point) hyperparameters of model #1, the Ensemble classification model.

Table 4. Minimum error hyperparameters of the Ensemble classification model

Hyperparameter	Value
Ensemble method	Bag
Number of learners	104
Maximum number of splits	1065
Number of predictors to sample	11

Fig. 9 shows the validation confusion matrix of the Ensemble classification model. We can observe that most cases were correctly predicted; however, the false discovery rate is higher in some categories, especially Leukemia with thrombocytopenia. Fig. 10 shows all categories' positive predictive values (PPV) and false discovery rates (FDR). This figure shows that the PPV for the category of Leukemia with thrombocytopenia is only 76.9%. The low value in Leukemia with thrombocytopenia might be because the dataset is imbalanced, containing only 11 observations for this category, which is only 0.86% of all observations. However, the PPV is satisfactory for all other categories, between 97.7% and 100%.

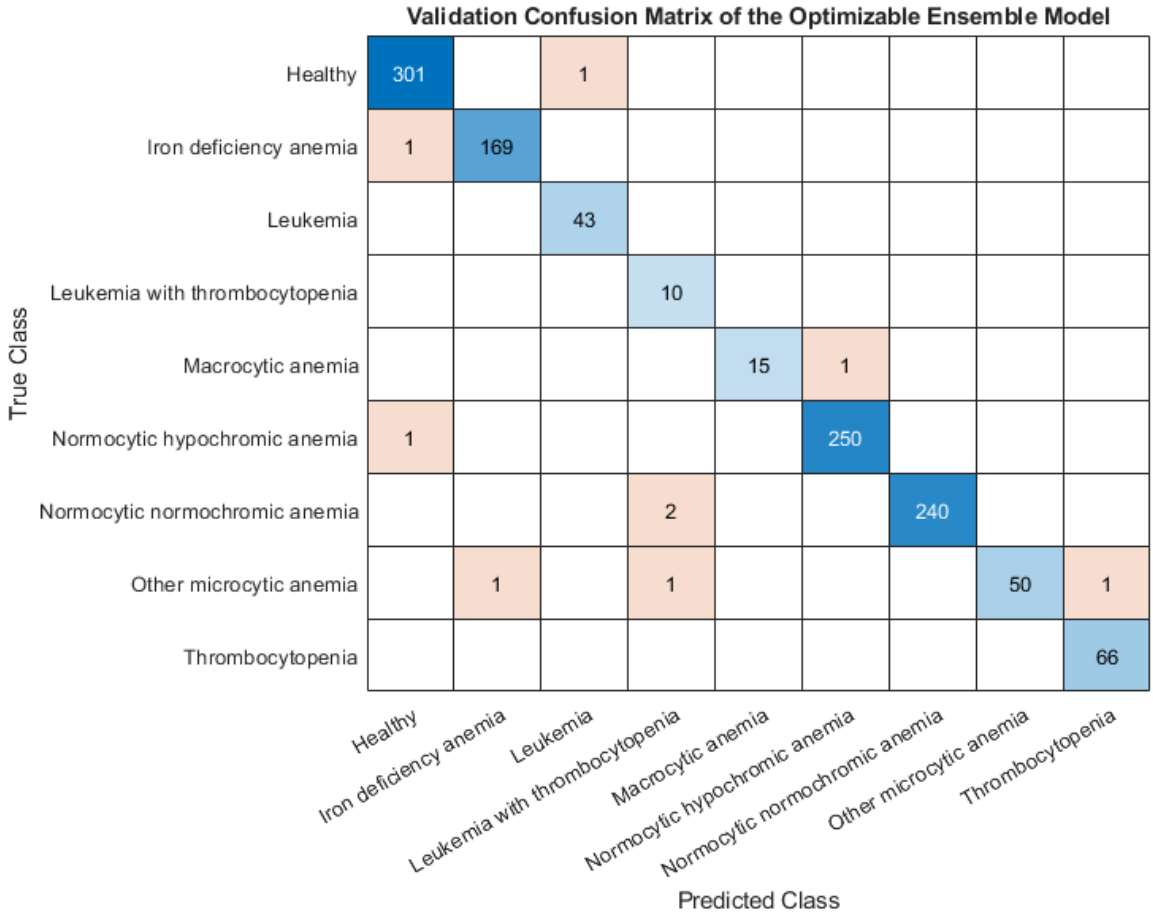

Fig. 9 Validation confusion matrix of the Ensemble classification model

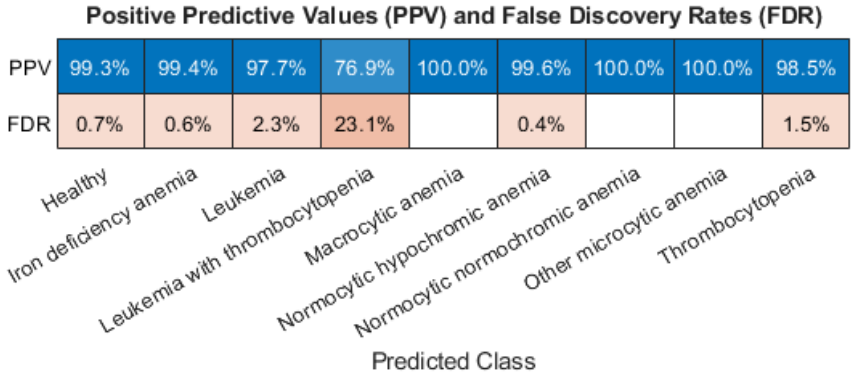

Fig. 10 Positive predictive values (PPV) and false discovery rates (FDR) in the validation confusion matrix of the Ensemble classification model

After training and validation, the final classification models were tested on unseen data (10% of observations). As we have seen in Table 3, the test accuracy was 100% for model #1, the Ensemble classification model. Fig. 11 shows the test confusion matrix of this model. This model correctly classified all observations in the test set.

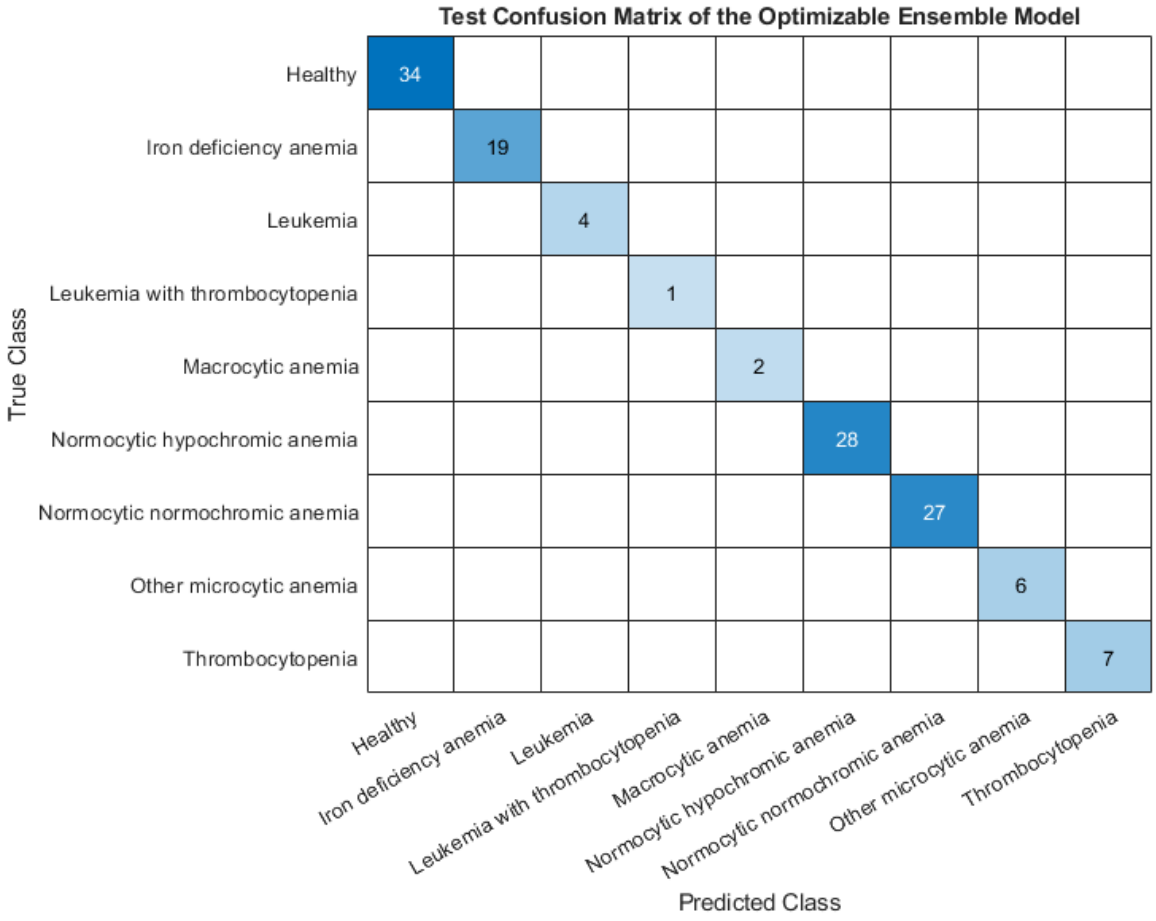

Fig. 11 Test confusion matrix of the Ensemble classification model

Finally, we wanted to determine how much each predictor contributed to the final Ensemble classification model. For this reason, we calculated the permutation feature importance. The results are shown in Fig. 12. As we can see on the chart, there were 6-7 important predictors; the most important feature was the amount of hemoglobin (HGB), followed by mean corpuscular volume (MCV), mean corpuscular hemoglobin concentration (MCHC), mean corpuscular hemoglobin (MCH), the number of platelets in the blood (PLT), the count of white blood cells (WBC), and the hematocrit test (HCT). The mean importance values of other predictors of the Ensemble classification model are very low; their values are near zero.

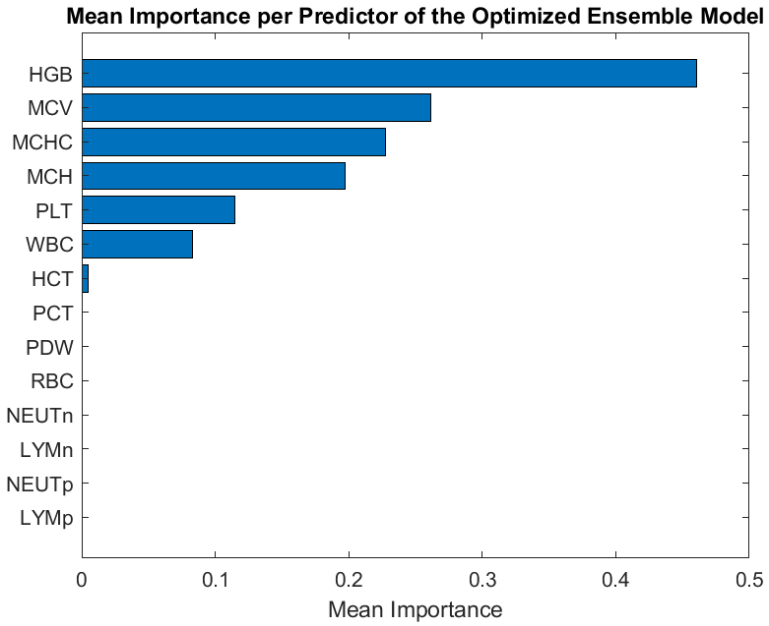

Fig. 12 Mean importance per predictor of the Ensemble classification model

IV. DISCUSSION

As we can see in the results, machine learning can help correctly diagnose amnesia and classify it into various types. However, finding the suitable classification model with the best accuracy is crucial. For this reason, we trained and tested nine optimizable classification models. The best model for our classification was the Ensemble model, where 99.22% validation and 100% test accuracy were reached. However, the positive prediction value (PPV) for one of the categories was lower (76.9%) than for other categories (97.7– 100%), probably because of the low number of observations for this category in the dataset. Furthermore, after calculating the permutation feature importance, we observed that the most important predictor of the Ensemble model was the amount of hemoglobin in the blood.

We can find similar results in the literature related to the anemia classification using the CBC data. Pullakhandam and McRoy [14] used a dataset containing about 20,000 samples to classify iron deficiency anemia from CBC data using machine learning. They trained and tested multiple machine learning algorithms, reaching 97% accuracy with Logistic Regression, Random Forest, KNN, Gradient Boosting, and XGBoost classification models.

Vohra et al. [15] also used CBC data to classify anemia using machine learning algorithms. The dataset contained 11 attributes and 364 observations. First, they used the original dataset on several classification models. The best accuracy was reached using the Logistic Regression classification model (94.44% accuracy using the hold-out method and 92.85% using the 10-fold cross-validation). Next, after feature selection, the accuracy increased; the best models in this case were the Decision Tree classification model (96.1% accuracy using the hold-out method) and the Multilayer Perceptron classification model (95.31% accuracy using the 10-fold cross-validation). Finally, they utilized the synthetic minority oversampling technique (SMOTE) to balance the dataset; the best model in this case was the Multilayer Perceptron classification model (99.35% accuracy using the hold-out method and 94.21% using the 10-fold cross-validation).

Yıldız et al. [13] used a dataset containing 1663 samples and 25 attributes to predict 12 different anemia types. They reached 85.6% accuracy using the Bagged Decision Trees classification model.

Even though our model's validation accuracy is 99.22%, we believe there are possibilities to improve the model. As we saw in the results, the PPV was low for one of the categories (76.9%) because of the imbalanced data. This could be improved by using more observations for the problematic category. If there is no way to get more real observations, generating synthetic data might help. After applying some

oversampling techniques [20], [21] and training the classification models with the modified dataset, the accuracy might improve for all categories. This could be part of future research.

V. CONCLUSION

After data exploration, we compared the accuracy of 9 optimizable classification models to diagnose anemia and classify them into eight anemia types. For our dataset, the best model was the Ensemble classification model using the bag ensemble method, which reached 99.22% validation accuracy and 100% test accuracy. Among other important predictors in the dataset, the most important feature for our Ensemble classification model was the amount of hemoglobin in the blood.

These results could be used in further research related to healthcare to diagnose anemia and its types automatically, time-savingly, less costly, and accurately from CBC data. The steps and methodology used in this paper can be applied to other datasets for anemia diagnosis and in other areas where it is needed to identify different categories from various data. Even though our best model was the Ensemble classification model, other machine learning models might achieve better results for other datasets. Training and testing different classification models to find the best, usually with the highest accuracy, that suits our needs is always recommended.

REFERENCES

[1] L. Végh, K. Czakóová, and O. Takáč, "Comparing Machine Learning Classification Models on a Loan Approval Prediction Dataset," *International Journal of Advanced Natural Sciences and Engineering Researches*, vol. 7, no. 9, pp. 98–103, Oct. 2023, doi: 10.59287/ijanser.1516.

[2] N. Annuš, "Usability of Artificial Intelligence to Create Predictive Models in Education," presented at the 15th International Conference on Education and New Learning Technologies, Palma, Spain, Jul. 2023, pp. 5061–5065. doi: 10.21125/edulearn.2023.1328.

[3] S. Szénási, G. Légrádi, and B. Vígh, "Machine Learning-Assisted Approach for Optimizing Step Size of Hill Climbing Algorithm," in *2024 IEEE 18th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, Timisoara, Romania: IEEE, May 2024, pp. 000425–000430. doi: 10.1109/SACI60582.2024.10619891.

[4] J. Udvaros and N. Forman, "Artificial Intelligence and Education 4.0," presented at the 17th International Technology, Education and Development Conference, Valencia, Spain, Mar. 2023, pp. 6309–6317. doi: 10.21125/inted.2023.1670.

[5] K. A. Tran, O. Kondrashova, A. Bradley, E. D. Williams, J. V. Pearson, and N. Waddell, "Deep learning in cancer diagnosis, prognosis and treatment selection," *Genome Medicine*, vol. 13, no. 1, p. 152, Sep. 2021, doi: 10.1186/s13073-021-00968-x.

[6] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," *Procedia Computer Science*, vol. 165, pp. 292–299, Jan. 2019, doi: 10.1016/j.procs.2020.01.047.

[7] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthc Technol Lett*, vol. 10, no. 1–2, pp. 1–10, Dec. 2022, doi: 10.1049/htl2.12039.

[8] L. Végh, O. Takáč, K. Czakóová, D. Dancsa, and M. Nagy, "Comparative Analysis of Machine Learning Classification Models in Predicting Cardiovascular Disease," *International Journal of Advanced Natural Sciences and Engineering Researches*, vol. 8, no. 6, pp. 23–31, Jul. 2024.

[9] S. Subramani *et al*., "Frontiers | Cardiovascular diseases prediction by machine learning incorporation with deep learning", doi: 10.3389/fmed.2023.1150933.

[10] N. B. Bahadure, R. Khomane, and A. Nittala, "Anemia Detection and Classification from Blood Samples Using Data Analysis and Deep Learning," *Automatika*, vol. 65, no. 3, pp. 1163–1176, Jul. 2024, doi: 10.1080/00051144.2024.2352317.

[11] A. Kovacevic, A. Lakota, L. Kuka, E. Becic, A. Smajovic, and L. G. Pokvic, "Application of Artificial Intelligence in Diagnosis and Classification of Anemia," in *2022 11th Mediterranean Conference on Embedded Computing (MECO)*, Budva, Montenegro: IEEE, Jun. 2022, pp. 1–4. doi: 10.1109/MECO55406.2022.9797180.

[12] P. T. Dalvi and N. Vernekar, "Anemia Detection Using Ensemble Learning Techniques and Statistical Models," in *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, Bangalore, India: IEEE, May 2016, pp. 1747–1751. doi: 10.1109/RTEICT.2016.7808133.

[13] T. Karagül Yıldız, N. Yurtay, and B. Öneç, "Classifying Anemia Types Using Artificial Learning Methods," *Engineering Science and Technology, an International Journal*, vol. 24, no. 1, pp. 50–70, Feb. 2021, doi: 10.1016/j.jestch.2020.12.003.

[14] S. Pullakhandam and S. McRoy, "Classification and Explanation of Iron Deficiency Anemia from Complete Blood Count Data Using Machine Learning," *BioMedInformatics*, vol. 4, no. 1, Art. no. 1, Mar. 2024, doi: 10.3390/biomedinformatics4010036.

[15] 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, p. e0269685, Jul. 2022, doi: 10.1371/journal.pone.0269685.

[16] “Anemia Types Classification.” Accessed: Aug. 14, 2024. [Online]. Available: https://www.kaggle.com/datasets/ehababoelnaga/anemia-types-classification
[17] “SPSS Software | IBM.” Accessed: Aug. 15, 2024. [Online]. Available: https://www.ibm.com/spss
[18] “MATLAB.” Accessed: Aug. 14, 2024. [Online]. Available: https://www.mathworks.com/products/matlab.html
[19] “Train models to classify data using supervised machine learning - MATLAB.” Accessed: Aug. 15, 2024. [Online]. Available: https://www.mathworks.com/help/stats/classificationlearner-app.html
[20] A. Viloria, O. B. Pineda Lezama, and N. Mercado-Caruzo, “Unbalanced data processing using oversampling: Machine Learning,” *Procedia Computer Science*, vol. 175, pp. 108–113, Jan. 2020, doi: 10.1016/j.procs.2020.07.018.
[21] R. Mohammed, J. Rawashdeh, and M. Abdullah, “Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results,” in *2020 11th International Conference on Information and Communication Systems (ICICS)*, Apr. 2020, pp. 243–248. doi: 10.1109/ICICS49469.2020.239556.