Using Machine Learning Models to Detect Anemia and Evaluate Accuracy Across Patient Subgroups in Blood Test Data

Rahini Chanda, Dougherty Valley High School, San Ramon, United States

ABSTRACT

Anemia is a widespread blood disorder currently affecting 2 billion people around the world. While machine learning (ML) has shown potential for enhanced detection of the condition using complete blood count (CBC) data, previous studies have found that ML models are unable to make consistent predictions across different demographic subgroups and patient populations. The study discussed in this paper trained and assessed two supervised ML models, Random Forest and Logistic Regression, on a CBC anemia dataset from the Eureka Diagnostic Center (EDC) in Lucknow, India, consisting of 364 patients. Hemoglobin (HGB) and features mathematically derived from it were excluded as an effort to prevent data leakage. During the study, Random Forest performed better than Logistic Regression in every accuracy metric considered for both overall testing and subgroup analysis, but both models performed poorly for elderly patients, aged 65 years and older. An ML framework, SHapley Additive exPlanations (SHAP), was used to analyze the importance of each of the features in the dataset, with red blood cell (RBC) count, packed cell volume (PCV), mean cell volume (MCV), and red cell distribution width (RDW) as the most important factors, while age and sex did not have substantial influence on the results. Both models were externally validated on a separate dataset from the Al-Zahraa Al-Ahly Hospital (AAH) in Iraq, which resulted in high accuracy and precision, but recall declined severely, suggesting that many true anemia cases were misdiagnosed when these ML models were applied to an unknown and different patient population. The results suggest that ML models could be used to detect anemia accurately within the population they were trained on, but for real-world use, the models will need to be trained on more diverse datasets.

KEYWORDS: Anemia; Machine Learning; Random Forest; Logistic Regression

INTRODUCTION

Anemia is a common blood disorder where the body does not have enough hemoglobin (HGB) or healthy red blood cells to carry enough oxygen to its tissues. Because of this, patients with anemia often feel shortness of breath, daily weakness, and fatigue (1). Anemia is an overarching term for many similar blood conditions, and there are over 400 types of anemia that are classified into several specific categories (2). The various types of anemia are microcytic anemias (e.g., iron deficiency anemia, anemia of chronic disease, thalassemia, etc.), normocytic anemias (e.g., early-stage iron deficiency, early-stage anemia of chronic disease, aplastic anemia, sickle cell disease, etc.), and macrocytic anemias (e.g., liver disease, alcoholism, vitamin B12 deficiency, etc.) (3). Approximately 2 billion people in the world have anemia, or in other words, approximately 1 in 4 people have an anemic condition. However, many patients are unaware that they have the condition, and many of those who are unaware may be severely at risk (4). If left untreated, anemia can lead to serious health issues such as arrhythmias (a condition involving an irregular heartbeat), an enlarged heart, or even heart failure. Thus, it is becoming increasingly important to better diagnose and treat anemia before patients' lives are

at severe risk (1). Additionally, anemia is more prevalent in minority races and ethnicities, including Black, Hispanic, and Asian patients. Therefore, improving its diagnosis and detection can help bridge ethnicity and research gaps in healthcare, improving inclusivity in the healthcare industry overall (5).

In the current age, artificial intelligence (AI) and machine learning (ML) are becoming increasingly important in the medical industry, as doctors can often overlook patterns in patient data and misdiagnose patients due to burnout. With that said, AI and ML should not be used solely on their own when handling patient data and making conclusions, but as a tool for clinicians and healthcare professionals to help better diagnose diseases/conditions and reduce the time needed (6).

Researchers around the world are currently applying AI and ML for disease detection to analyze their benefits and limitations. For example, a recent 2025 study, “Machine Learning-Based Models for Screening of Anemia and Leukemia Using Features of Complete Blood Count Reports,” was conducted in Islamabad, Pakistan, using this research approach. The researchers built ML models that were trained on complete blood count (CBC) features of a small patient dataset from Pakistan. Their models performed relatively well on the testing data (also from Pakistan), achieving an overall accuracy of about 97%-98%. However, the accuracy of their model decreased when it was tested on a completely different dataset from Iraq (overall accuracy of about 74%-75%), suggesting overfitting or other potential reasons based on differences in the demographics of the training and external data (7). Additionally, another study conducted in 2026, “Machine Learning-Driven Anemia Diagnosis: A Comparative Study Using Blood Biomarkers from Complete Blood Count Data,” trained and tested five different ML models to detect anemia, including Random Forest and Logistic Regression, on CBC blood biomarker data. Random Forest had the highest overall accuracy of the five ML models, and Logistic Regression had the second highest overall accuracy. Although Logistic Regression performed slightly lower than Random Forest in the study, it was noted for performing relatively well and being able to run relatively quickly, while being simple to interpret, which is important for clinical use (8).

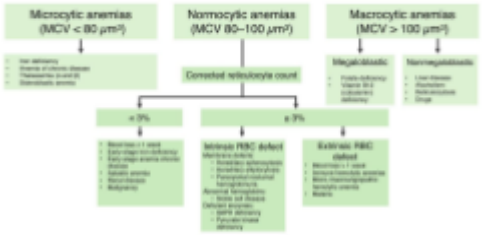

Figure 1. *Classification of anemia subtypes based on the mean corpuscular volume (MCV), including microcytic, normocytic, and macrocytic anemias and their underlying causes (3).*

The research discussed in this paper expands on the ideas of the previous studies by evaluating Random Forest and Logistic Regression separately across two patient subgroups, age and sex, instead of only reporting the overall accuracy. Age and sex were chosen as the subgroups for evaluation because normal HGB levels (corresponding to anemia) are known to differ by sex and change during a lifetime, making these subgroups valuable for testing model accuracy across demographics (9). The aim of this paper is to analyze whether the ML models, Random Forest and Logistic Regression, can accurately detect anemia from CBC data across different populations, and whether the accuracy remains consistent across different patient subgroups, such as age and sex.

METHODS AND MATERIALS

Dataset and Preprocessing

The models were originally trained and tested on an 80/20 split of a CBC dataset from the Eureka Diagnostic Center (EDC) in Lucknow, India, consisting of 364 patients (10). An anemia label was created using the HGB cutoffs for both males and females. For men, an HGB value less than 13.0 g/dL is classified as anemic, and for women, an HGB value less than 12.0 g/dL is classified as anemic, as per standardized clinical cutoffs (9). Binary values (0 and 1) were used in the dataset to encode the patient sex, but it was not specified which value corresponded to which sex. This was inferred through calculations based on accepted HGB values with males typically having higher HGB values than females. By comparing the mean and median HGB values between both sex groups, it was determined that the value 0 corresponded with male patients, while the value 1 corresponded with female patients: the 0 group had a higher mean (12.63 g/dL HGB) and median (13.1 g/dL HGB) than the 1 group's mean (10.99 g/dL) and median (11.1 g/dL HGB). Patients were also grouped into four different age groups for subgroup evaluation: pediatric (less than 18 years), young adult (18-39 years), middle age (40-64 years), and older adults (65 years and older). Before being used for training and testing, the dataset was cleaned by removing a nonessential row containing column labels in unabbreviated form, eliminating blank rows at the end of the dataset, and deleting extraneous whitespace. After the dataset was ready for use, it was determined that 208 patients were anemic (57.1%) and 156 patients were non-anemic (42.9%). This class distribution was consistent with the finding that anemia is substantially more prevalent within certain populations, as discussed earlier.

Figure 2. *Sample of raw patient data from the EDC CBC dataset prior to cleaning, showing the original column set and the unlabeled binary sex encoding of 0 and 1.*

Feature Selection and Leakage Prevention

The final features used in the training and testing of the models were age, sex, red blood cell (RBC) count, packed cell volume (PCV), mean cell volume (MCV), red cell distribution width (RDW), total leukocyte count (TLC), and platelet count (PLT). HGB was removed from the models' input features because these values directly defined the anemia label, and the models were making predictions based solely on HGB values (data leakage), causing unrealistically high accuracy results of 1.00 AUROC (Area Under the Receiver Operating Characteristic). Including HGB allowed the models to only reference the values that were used in creating the anemia label, preventing the models from learning the actual patterns in the dataset. Similarly, mean corpuscular hemoglobin (MCH) and mean corpuscular hemoglobin concentration (MCHC) were removed from the dataset to further prevent data leakage, as both are mathematically derived from HGB: MCH = HGBx10/RBC and MCHC =

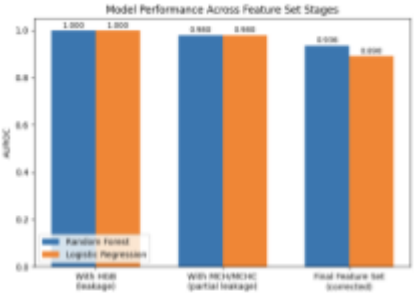

Figure 3. *Random Forest and Logistic Regression AUROC values across three feature configurations, showing the impact of data leakage from HGB, MCH, and MCHC.*

HGBx100/PCV. This change was made as the models were still producing unrealistic results with an AUROC of 0.98. The Islamabad study discussed earlier did not look this closely at the derived features to find data leakages, which may be a potential limitation explaining why their models didn't perform well on external data (7).

Model Development

Two supervised ML models, Random Forest and Logistic Regression, were implemented in this study (11). The EDC dataset was split into 80% training (291 patients) and 20% testing (73 patients) sets. Stratified sampling was used to maintain the class balance of 57% anemic patients and 43% non-anemic patients in both the training and testing sets. A 5-fold cross-validation was performed on the training set: the training data was split into five groups to verify model consistency. The features were scaled using the StandardScaler for the Logistic Regression model because of its sensitivity to large differences in feature scale; the platelet count values were in the hundreds while the mean cell volume values were in the tens. This step was not required for Random Forest because, as a tree-based model, the data is split based on thresholds instead of distances between values. Both the Logistic Regression and Random Forest models were trained using a fixed random state (i.e., random state = 42) to ensure consistent results. 100 estimators with default max depth were used to train Random Forest, while a maximum of 1000 iterations were used to train Logistic Regression.

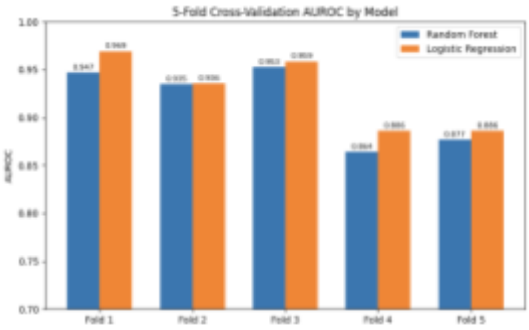

Figure 4. *Random Forest and Logistic Regression AUROC scores across each of the five cross-validation folds, showing model consistency before using the test set.*

External Validation

An independent CBC dataset from the Al-Zahraa Al-Ahly Hospital (AAH) in Iraq (12) was used to externally validate and test the Random Forest and Logistic Regression models, which were both trained on the EDC dataset. This was done to test whether the trained models could predict anemia accurately on a dataset of an unknown population. Before being used, the AAH dataset was cleaned by removing data of patients with implausible values (i.e., very large numbers, negative numbers, etc.), which reduced the dataset from 500 to 477 patients. Since the AAH dataset did not contain age or sex columns, the Random Forest and Logistic Regression models were re-trained on the EDC dataset using only the features that were safe from data leakage and available in both the EDC and AAH datasets: RBC, PCV, MCV, RDW, TLC, and PLT. Also, because the AAH dataset did not contain a sex column, an HGB cutoff of 13.0 g/dL was used to define the anemia label instead of specific cutoffs based on sex that were previously applied to the EDC dataset. The models were not exposed to any AAH data during training to ensure that the AAH data was a genuine test of the generalization of ML models on different populations. Additionally, it is important to

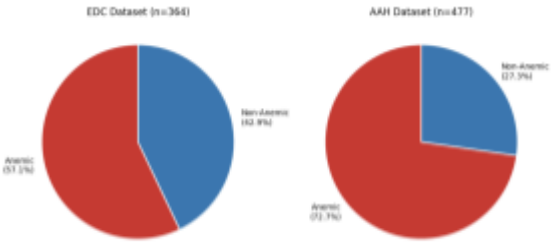

Figure 5. *Comparisons of anemic and non-anemic class distribution between the EDC dataset and the AAH dataset, showing the class balance difference between the two populations.*

note that the AAH dataset had a different class distribution than the EDC dataset: 72.7% anemic vs. 27.3% non-anemic (AAH) compared to 57.1% anemic vs. 42.9% non-anemic (EDC). This is a limitation when comparing the models' performance on each of the datasets.

RESULTS

Overall Model Performance

Both the Random Forest and Logistic Regression models were tested on the remaining 20% of the EDC dataset (80% was used for training, as discussed previously). During testing, Random Forest resulted in an overall AUROC of 0.936, precision (TP/[TP+FP]) of 0.886, recall (TP/[TP + FN]) of 0.929, and F1 score (2xPxR/[P+R]) of 0.907. Logistic Regression resulted in an AUROC of 0.890, precision of 0.841, recall of 0.881, and F1 score of 0.860. Based on these results, Random Forest performed better than Logistic Regression on every metric measuring the accuracy. It is also important to note that the recall score was the most important metric as it is the most important in clinical screening to avoid missing true positive cases.

```
=== Random Forest (Test Set) ===
AUROC: 0.9362519201228879
Precision: 0.8863636363636364
Recall: 0.9285714285714286
F1: 0.9069767441860465

=== Logistic Regression (Test Set) ===
AUROC: 0.890168970814132
Precision: 0.8409090909090909
Recall: 0.8809523809523809
F1: 0.8604651162790697
```

Figure 6. *Test set evaluation results for the Random Forest and Logistic Regression models on the EDC dataset.*

Subgroup Performance: Sex and Age

When the models were evaluated by sex, Random Forest resulted in an AUROC of 0.919 for male patients and an AUROC of 0.970 for female patients, while Logistic Regression resulted in an AUROC of 0.895 for male patients and 0.910 for female patients. Additionally, Logistic Regression also resulted in a recall of 1.000 for male patients, meaning that every anemic male patient was correctly classified; however, this was at the cost of a lower precision of 0.762. For female patients, Logistic Regression's recall was 0.808, indicating that several anemic females were misclassified when compared to males. When the models were evaluated by age group, both Random Forest and Logistic Regression models performed poorly on patients in the older adults subgroup (patients 65 years and older). For the older adults group, Random Forest resulted in an AUROC of 0.889 and Logistic Regression resulted in an AUROC of 0.815. For the young adults group (patients 18-39 years), Random Forest resulted in an AUROC of 0.983 and Logistic Regression resulted in an AUROC of 0.903. The pediatric group resulted in an AUROC of 1.000 for both models, but this group consisted of only three patients in the test set, making the result inconclusive and unreliable. Overall, Random Forest

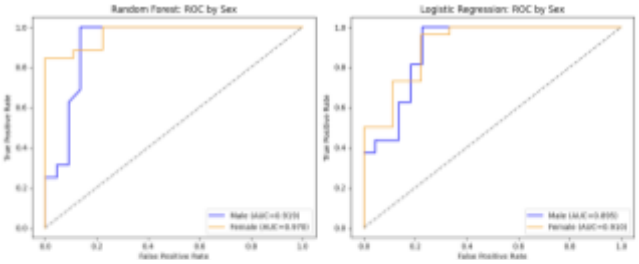

Figure 7. *Random Forest and Logistic Regression ROC curves by sex on the EDC dataset.*

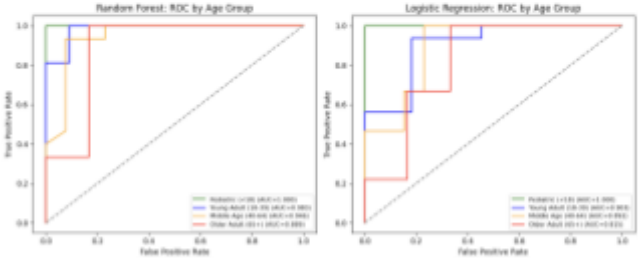

Figure 8. *Random Forest and Logistic Regression ROC curves by age group on the EDC dataset.*

performed better than Logistic Regression in both the subgroups evaluated, suggesting it is more consistent in anemia detection across different patient demographics.

Feature Importance Analysis

SHapley Additive exPlanations (SHAP) was used to analyze features that influenced each model's predictions the most (13). It was found that RBC, PCV, MCV, and RDW were consistently ranked as the most important features for both models. Age and sex did not impact the model predictions as much as the CBC features, suggesting that the models did not singularly depend on demographic information in anemia prediction. The feature importance rankings were mostly consistent across both the sex and age subgroups, indicating that differences in model performance likely come from how the data is spread within each group, rather than the models having different reasoning for different patients.

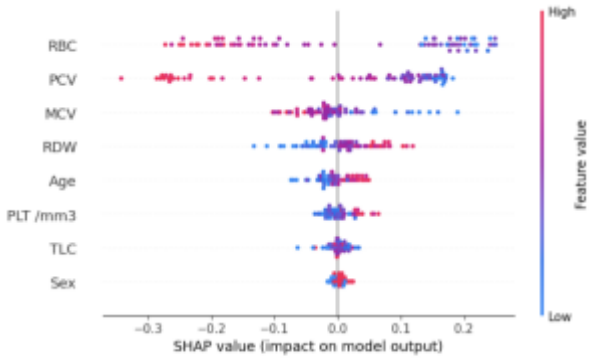

Figure 9. *Example of a SHAP summary plot for the Random Forest model, showing feature importance.*

External Validation

The initial results after testing with the AAH dataset were promising: Random Forest resulted in an AUROC of 0.953 and Logistic Regression resulted in an AUROC of 0.968, with both models resulting in an approximate precision of 0.984. However, recall decreased noticeably for both models in external validation: Random Forest resulted in a recall of 0.697 and Logistic Regression resulted in a recall of 0.712, meaning that about 3 in 10 truly anemic patients in the AAH data were being incorrectly classified by the models as non-anemic. This suggests that while both Random Forest and Logistic Regression models can classify several anemic patients, they are noticeably less effective when being used to classify patients of a different population than the population they were originally trained on.

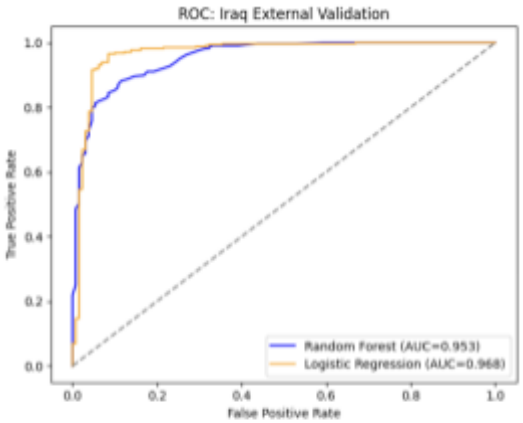

Figure 10. *Random Forest and Logistic Regression ROC curves for external validation on the AAH dataset.*

DISCUSSION

In overall evaluation, subgroup analysis, and external validation, Random Forest performed better than Logistic Regression. However, both models struggled with the older adults group, regardless. The results suggest that a model can predict anemia in patients relatively well while still performing inconsistently across specific subgroups, meaning that there is a difference between model accuracy and model fairness. This difference is especially

important in clinical screening, where recall is one of the most important metrics, as missing a true anemia case (resulting in a false negative) has greater consequences than a false positive. Therefore, one of the most significant results of this study is the decline in recall during external validation on the AAH dataset, as this indicates that both the Random Forest model and the Logistic Regression model would miss a large proportion of true anemia cases if used on an unknown patient population. It is also important to note that the EDC test set consisted of only 73 patients, suggesting that a few misclassifications could largely affect the precision and recall values reported. Moreover, since there were no age and sex columns in the AAH dataset, the comparisons between the two datasets are limited because a general HGB value was used to create the anemia label for the AAH dataset, instead of the sex-specific HGB values used for the EDC dataset.

SUMMARY AND CONCLUSION

This study evaluated two supervised ML models: Random Forest and Logistic Regression. They were both trained on CBC data to detect anemia and were evaluated on overall performance, subgroup (age and sex) performance, and external validation. Throughout the study, Random Forest consistently performed better than Logistic Regression, although both models were able to accurately detect anemia overall. However, it is important to note that recall declined greatly when the AAH dataset was used for external validation. The decline suggests that both models would benefit and improve their results if trained on more diverse datasets containing information of different patient populations. With further training and testing on diverse patient data, ML models could be used to help healthcare professionals detect anemia earlier and more consistently, although they are not ready to be deployed for real-world use yet.

CONFLICT OF INTEREST: The author declares no conflicts of interest related to this work.

REFERENCES

1. Mayo Clinic Staff. (2026, May 5). *Anemia*. Mayo Clinic. Retrieved July 24, 2026, from https://www.mayoclinic.org/diseases-conditions/anemia/symptoms-causes/syc-20351360
2. Penn Medicine. (2026). *Anemia*. Penn Medicine. Retrieved July 24, 2026, from https://www.pennmedicine.org/conditions/anemia
3. Lecturio. (2024, December 17). *Anemia: Overview and Types*. Lecturio. Retrieved July 24, 2026, from https://www.lecturio.com/concepts/anemia-overview/
4. Ritchie, H. (2024, November 25). *Billions of people suffer from anemia, but there are cheap ways to reduce this*. Our World in Data. Retrieved July 24, 2026, from https://ourworldindata.org/billions-people-suffer-anemia-cheap-ways-reduce
5. Williams, A. M., Ansai, N., Ahluwalia, N., & Nguyen, D. T. (2024, December). *Anemia Prevalence: United States, August 2021–August 2023*. NCHS Data Briefs. Retrieved July 24, 2026, from https://www.ncbi.nlm.nih.gov/books/NBK612586/
6. Al-Antari, M. A. (2023). Artificial Intelligence for Medical Diagnostics—Existing and Future AI Technology! *Diagnostics*, *13*(4). https://pmc.ncbi.nlm.nih.gov/articles/PMC9955430/
7. Amjad, H., Hussain, Z., Hasan, M., & Hassan, M. U. (2025). Machine learning-based models for screening of anemia and leukemia using features of complete blood count reports. *Scientific Reports*. https://www.nature.com/articles/s41598-025-21279-w

8. Xhepaliu, A., Adar, N., & Leka, M. (2026). *Machine Learning-Driven Anemia Diagnosis: A Comparative Study Using Blood Biomarkers from Complete Blood Count Data*. SpringerLink. Retrieved July 24, 2026, from https://link.springer.com/chapter/10.1007/978-3-032-07373-0_30
9. Addo, O. Y., Williams, A., Young, M. F., Sharma, A. J., Mei, Z., Kassebaum, N. J., Socorro Jeffereds, M. E., Suchdev, P. S., & Yu, E. X. (2021). *Evaluation of Hemoglobin Cutoff Levels to Define Anemia Among Healthy Individuals*. ResearchGate. Retrieved July 24, 2026, from https://www.researchgate.net/publication/353743396_Evaluation_of_Hemoglobin_Cutoff_Levels_to_Define_Anemia_Among_Healthy_Individuals
10. Vohra, R., Pahareeya, J., & Hussain, A. (2021, April 22). *Complete Blood Count Anemia Diagnosis* [Data set]. Mendeley Data. https://doi.org/10.17632/dy9mfjchm7.1
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011, October). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research. Retrieved July 24, 2026, from https://www.jmlr.org/papers/v12/pedregosa11a.html
12. Sami, S., & Farhan, A. (2022, November 21). *CBC Dataset* [Data set]. Mendeley Data. https://doi.org/10.17632/28s2bhdjfd.1
13. Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. Neural Information Processing Systems. Retrieved July 24, 2026, from https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html