BioMedInformatics

MDPI

Article

Classification and Explanation of Iron Deficiency Anemia from Complete Blood Count Data Using Machine Learning

Siddartha Pullakhandam and Susan McRoy *

Department of Computer Science, University of Wisconsin-Milwaukee, Milwaukee, WI 53211, USA; pullakh2@uwm.edu

* Correspondence: mcroy@uwm.edu; Tel.: +1-414-229-6695

Abstract: Background: Currently, discriminating Iron Deficiency Anemia (IDA) from other anemia requires an expensive test (serum ferritin). Complete Blood Count (CBC) tests are less costly and more widely available. Machine learning models have not yet been applied to discriminating IDA but do well for similar tasks. Methods: We constructed multiple machine learning methods to classify IDA from CBC data using a US NHANES dataset of over 19,000 instances, calculating accuracy, precision, recall, and precision AUC (PR AUC). We validated the results using an unseen dataset from Kenya, using the same model. We calculated ranked feature importance to explain the global behavior of the model. Results: Our model classifies IDA with a PR AUC of 0.87 and recall/sensitivity of 0.98 and 0.89 for the original dataset and an unseen Kenya dataset, respectively. The explanations indicate that low blood level of hemoglobin, higher age, and higher Red Blood Cell distribution width were most critical. We also found that optimization made only minor changes to the explanations and that the features used remained consistent with professional practice. Conclusions: The overall high performance and consistency of the results suggest that the approach would be acceptable to health professionals and would support enhancements to current automated CBC analyzers.

Keywords: explainable AI; biomedical data; blood disorders; anemia; ferritin; iron deficiency; iron deficiency anemia; machine learning; SHAP

Citation: Pullakhandam, S.; McRoy, S. Classification and Explanation of Iron Deficiency Anemia from Complete Blood Count Data Using Machine Learning. *BioMedInformatics* **2024**, *4*, 661–673. https://doi.org/10.3390/biomedinformatics4010036

Academic Editors: Pentti Nieminen and Jörn Lötsch

Received: 12 January 2024
Revised: 15 February 2024
Accepted: 22 February 2024
Published: 1 March 2024

1. Introduction

It is currently unknown how well data from routine Complete Blood Count (CBC) tests alone can be used to identify a common form of anemia (iron deficiency), potentially replacing a more expensive test (serum ferritin) now used by experts for diagnosis. Here, we address this question through a novel application of Machine Learning methods involving model training and feature selection. We then apply methods for explaining the output of Machine Learning models to reveal which aspects of the CBC results contribute most to a diagnosis and to what extent methods meant to improve model performance impact the role of different features. These additional tests allow us to confirm that the machine learning model's "reasoning" is consistent with clinical judgment. This use of explanatory AI to show the impact of feature selection on feature importance is also unique to this study.

Anemia, defined by blood hemoglobin levels lower than specific age and gender reference levels [1], is of global health concern affecting the health and productivity of populations [2]. The 2023 Global Burden of Disease study estimated that the global prevalence of anemia is about 24.5%, corresponding to about 1.98 billion people [2]. Iron deficiency is a common, yet potentially preventable, cause of anemia because iron is an essential nutrient for both the synthesis and functioning of hemoglobin [3]. Insufficient consumption or malabsorption of dietary iron results in the reduced synthesis of red blood cells and blood hemoglobin levels [1]. Thus, supplementation of iron is routinely practiced, improving blood hemoglobin levels in populations where the prevalence of anemia is more than 40% [1]. Although the most common cause of anemia has been iron deficiency, the

deficiencies of other nutrients such as vitamin B12 (cyanocobalamin), vitamin B9 (folic acid), genetic disorders of hemoglobin, and tropical diseases also contribute to anemia prevalence [2,4]. Therefore, in addition to hemoglobin, the measurement of serum ferritin, an indicator of body iron stores, is required to identify the cause of anemia, which can then be treated with iron supplements [5]. However, the measurement of serum ferritin requires additional blood processing and advanced immunological testing [5], which may not be practical in many hospital/public health settings with limited resources.

Although serum ferritin is widely used to identify the cause of anemia [5], several reports have suggested that Red Blood Cell (RBC) indices such as Red blood cell Distribution Width (RDW), Mean Corpuscular Volume (MCV), and hemoglobin can be used as a differential diagnostic tool for the identification of iron deficiency anemia [6–9]. The current standard complete blood count (CBC) lab test, which is performed using automated hematology analyzers, measures blood hemoglobin and multiple blood cell indices such as mean corpuscular volume (MCV), packed cell volume (PCV), hematocrit (HCT), red blood cell number (RBCs), red blood cell distribution width (RDW), platelet (PLT) count, and data on total and differential white blood cells (WBC), neutrophils, and monocytes [6–9]. Therefore, it is theoretically possible to classify the type of anemia (whether it is due to iron deficiency or not) based on CBC data alone. If this approach is successful and generalizable, the model could eliminate the need for serum ferritin laboratory tests for IDA, and the associated costs.

Machine learning (ML) algorithms are increasingly being used in medicine for the classification of diseases, predicting the clinical outcome [10,11]. Indeed, many studies have attempted to diagnose or classify anemia based on blood cell variables [12,13], demographic variables [14,15], images of palm [16], conjunctiva, or fingertips [17–19], and sickle cell anemia from images of blood smears [20], but all these studies were to diagnose anemia, rather than IDA specifically. Some studies also reported differential diagnosis of IDA and ß-thalassemia with high accuracy [12,21,22]. In the few other studies that include IDA classification, either serum iron parameters or serum ferritin was used as features along with CBC data [23–27]. However, a recent study reported the prediction of low ferritin levels or IDA among adult anemic subjects (more than 18 years of age) in referred lab tests with 90–98% specificity and sensitivity, using a random forest algorithm [28]. However, as suggested in a recent review [29], these models need to be validated in larger datasets and across larger age and gender subgroups to show robust generalizability. No prior studies have been shown to discriminate IDA effectively for a wide sample of subjects using only features from the CBC alone. (See Supplementary Table S1 for more information.) More importantly, in a medical setting, explaining the risk factors (features) that contributed to a diagnosis of disease is of utmost importance to guide the clinician to make informed decisions (and to trust the diagnosis if it was not anticipated). Explainable AI (XAI) algorithms provide a useful interpretation of individual feature contributions to the diagnostic model [30].

For this study, we used publicly available pooled survey data from the National Health and Nutrition Examination Surveys NHANES (2003–2020) (https://wwwn.cdc.gov/nchs/nhanes/Default.aspx accessed on 21 February 2024) conducted by the US Centers of Disease Control and Prevention (CDC) [31], to test the performance of machine learning models in discriminating between IDA and Non-IDA, where Non-IDA includes other causes of anemia or no anemia. Further, we also handled the class imbalance (4.9% IDA (n = 972) vs. 95.1% non-IDA (n = 19,203) and analyzed the feature contributions to the model. Finally, we tested the generalizability of the model by validating it with data collected in a different setting.

2. Materials and Methods

2.1. Data Source

The publicly available NHANES data of 2003–2004, 2005–2006, 2007–2008, 2009–2010, 2015–2016, and 2017–2020 pre pandemic data were used for this analysis, where CBC,

serum ferritin, and demographic data are available. We examined data that included serum ferritin, as we used it to determine a gold standard label for training and testing. (For the survey between the years 2011 and 2013, serum ferritin data were not available, so those data were excluded). After removing instances that did not meet the study criteria, as described below, about 20,000 instances were available.

The demographic laboratory data (CBC and serum ferritin) were downloaded from the NHANES (https://wwwn.cdc.gov/nchs/nhanes/Default.aspx accessed on 21 February 2024), as .XPT files, transferred to a Jupiter notebook (v. 6.5.4), and combined using the pandas library, which implements machine learning algorithms in Python (v. 3.10.11) [32]. To preprocess the data, first the data were concatenated (all files of demography, CBC, and serum ferritin), and then each file was merged based on SEQUENCE IDs (SEQN) into a single data frame, to align the paired data of survey subjects across all the variables. The results were then converted into a .csv file for further analysis. The demography, CBC, and serum ferritin variable identifiers (column/variable names), units and their description are given in Table 1.

Table 1. Names and description of variables' units of demography, CBP, and serum ferritin variables.

Variable	Description	Units	Range (Min, Max)
RIDAGEYR	Age of the subject at the time of survey	Years	1 to 60 years
RIAGENDR	Gender of the subject	Male 1, Female 2	None
RIDEXPRG	Pregnancy status of the subject	1 for positive, 0 for negative	None
LBXWBCSI	White Blood Cell count	1000 cells/L	1.4, 23.4
LBXLYPCT	Lymphocytes percentage	%	4.2, 84.1
LBXMOPCT	Monocytes percentage	%	0.7, 40.4
LBXNEPCT	Segmented neutrophils percentage	%	2.4, 92.3
LBXEOPCT	Eosinophils percentage	%	5.4×10^{-79}, 34.1
LBXBAPCT	Basophils percentage	%	5.4×10^{-79}, 19.7
LBDLYMNO	Lymphocyte count	1000 cells/µL	0.2, 12.4
LBDMONO	Monocyte count	1000 cells/µL	5.4×10^{-79}, 3.8
LBDNENO	Neutrophil count	1000 cells/µL	0.2, 16.3
LBDEONO	Eosinophil count	1000 cells/µL	5.4×10^{-79}, 4.5
LBDBANO	Basophil count	1000 cells/µL	5.4×10^{-79}, 1.7
LBXRBCSI	Red Blood Cell count	10^6 cells/µL	2.61, 7.33
LBXHGB	Hemoglobin concentration	g/dL	6.1, 18.1
LBXHCT	Hematocrit	%	20.5, 54.9
LBXMCVSI	Mean Corpuscular Volume	fL	35.4, 116.8
LBXMC	Mean Corpuscular Hemoglobin Concentration	g/dL	25.2, 43.3
LBXMCHSI	Mean Corpuscular Hemoglobin	Pg	10.2, 56.2
LBXRDW	Red Cell Distribution Width	%	6.3, 36.5
LBXPLTSI	Platelet count	1000/µL	4, 1021
LBXMPSI	Mean Platelet Volume	fL	5, 13.5
LBXFER	Serum ferritin	µg/L	1.04, 200

Pg = Pecogram (10^{-12} g); µL = Microliter (10^{-3} mL); fL = femtoliter (10^{-12} mL); dL = Deciliter (100 mL).

2.2. Variable Selection and Data Cleaning

Except subject ID (SEQN, required for pairing with other datasets), age (RIDAGEYR), gender (RIAGENDR), and pregnancy (RIDEXPRG) data (required for the classification of anemia and iron deficiency), all other demographic variables were removed. The variables, LBXLYPCT, LBXMOPCT, LBXNEPCT, LBXEOPCT, LBXBAPCT, were also removed as their absolute counts were available in other variables.

The sequential exclusion of data and associated sample loss are given in Figure 1. Briefly, rows wherever LBXHGB or LBXFER is null were also removed, as both these features are necessary for IDA classification. Next any row with ferritin values of 150 µg/L for females, 200µg/L for males, which indicates iron overload [33], and any row with null values in any of the variables was removed to obtain paired data of all variables. The

RIDEXPRG is coded as 1 for positive pregnancy, and 2 and 3 for uncertain or could not ascertain the status at the time of the NHANES survey. Therefore, only 1 is considered positive pregnancy while all others are considered non-pregnant (coded as 0). The data of subjects above the age of 60 were excluded, as hemoglobin concentration at this age reduces, independent of iron status due to age-related physiology [34]. The total available paired data of all required demography, CBP, and serum ferritin data were 19,975.

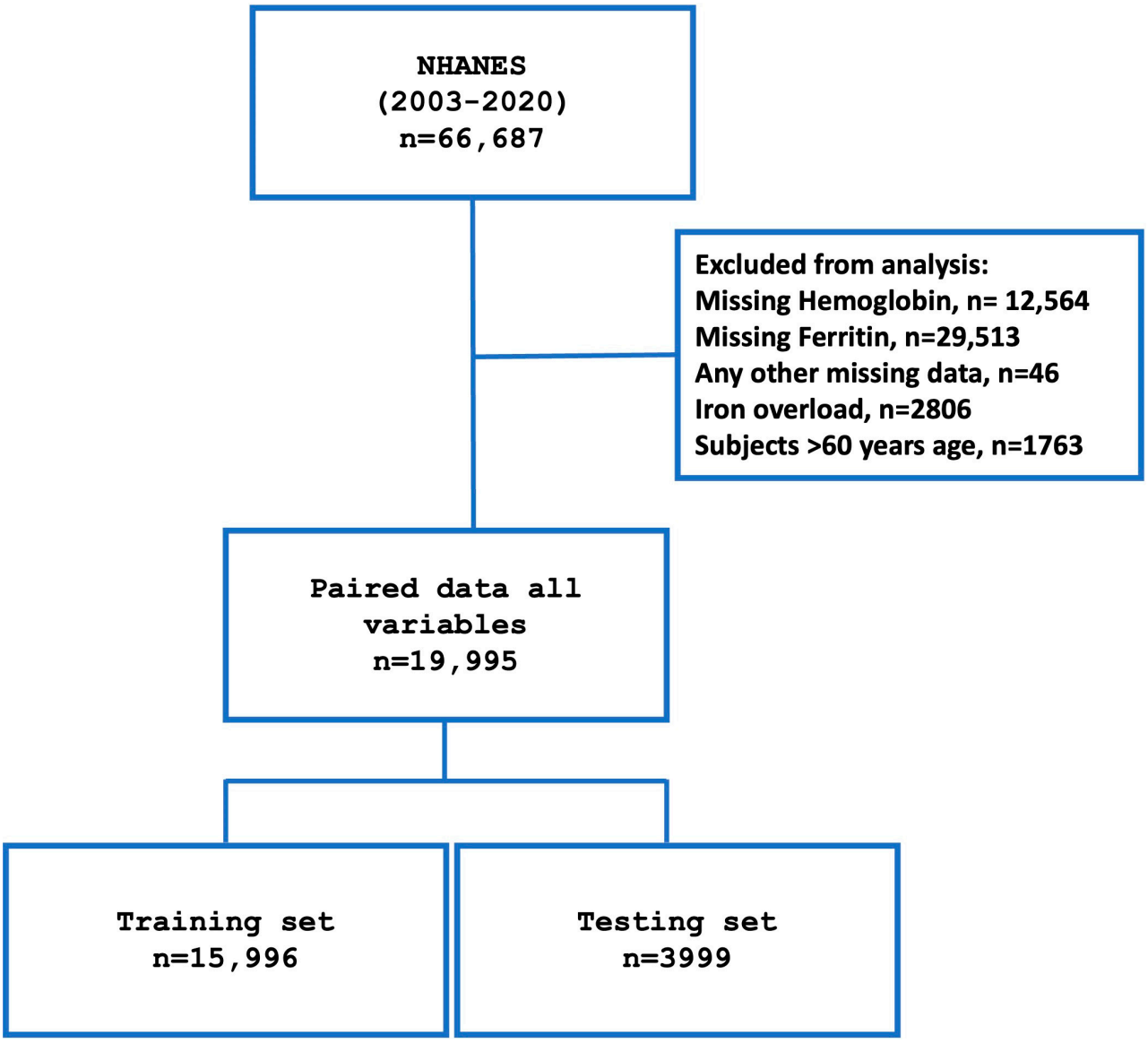

Figure 1. This flow diagram shows sequential data exclusion based on specific criteria and data splitting for training and testing for machine learning models.

2.3. Validation Dataset

To test the generalizability of the model in an independent different dataset, we tested the trained model's performance on an unseen dataset with the same features. The dataset used for cross validation was from a healthy adult Kenyan population (n = 502, and is publicly available [35,36]. Some instances (n = 26) were excluded, where age >60 years old; in the data used, age ranged from 18–60, and the gender distribution was 52.7% female. Since this study excluded subjects who were pregnant, for all samples of RIDEXPRG the value was considered zero. The proportions of anemia, ID, and IDA were 4.5%, 15.5%, and 3.78%, respectively.

2.4. Classification of Anemia, Iron Deficiency (ID), and Iron Deficiency Anemia (IDA)

The age, gender, pregnancy status, specific hemoglobin (LBXHGB), and serum ferritin (LBXFER) reference values (Table 2) suggested by the WHO were used for the classification of anemia and ID [1,33]. A subject who was both positive for anemia and ID was classified as IDA, and otherwise was considered as non-IDA.

Table 2. World Health Organization (WHO) reference values for classification of anemia, iron deficiency (ID).

Age	Biological Gender	Hemoglobin Reference (g/100 mL) for Anemia	Serum Ferritin Reference for Iron Deficiency (μg/L)
less than 5 years	Any	11	12
5–11 years	Any	11.5	15
12–14 years	Any	12	15
15 years and above	Male	13	15
15 years and above	Non pregnant female	12	15
	Pregnant female	11	

2.5. Data Aggregation and Preprocessing

The proportion of anemia, ID, and IDA or mean of numerical variables along with their 95% confidence intervals (Cis) were computed using the appropriate "group by" function in the stats package in the SciPy library. The bar graphs with 95% CI error bars for all numerical variables stratified by IDA status were generated using matplotlib software, and the 95% CIs were non-overlapping and hence considered statistically significant ($p < 0.05$). Since the units for different features varied, all the data were normalized using the StandardScaler method from the scikit-learn Library prior to the data input in ML models.

2.6. Classification of IDA by Machine Learning (ML) Algorithms

The full pipeline for training and validation is shown in Figure 2. The classification of IDA and Non-IDA was initially tested using multiple ML algorithms, namely logistic regression (LR), random forest (RF), K-Nearest Neighbors (KNN), Naïve Bayes (NB), gradient boosting (GB), and XGBoost (XGB), with all the features except SEQN and LBXFER variables. (These algorithms form a representative set of linear and nonlinear classifiers available in scikit-learn and are suitable for datasets with thousands, but not millions, of instances. They require no tuning of hyperparameters.) The target feature of IDA is binary coded, with IDA as 1 and non-IDA as 0.

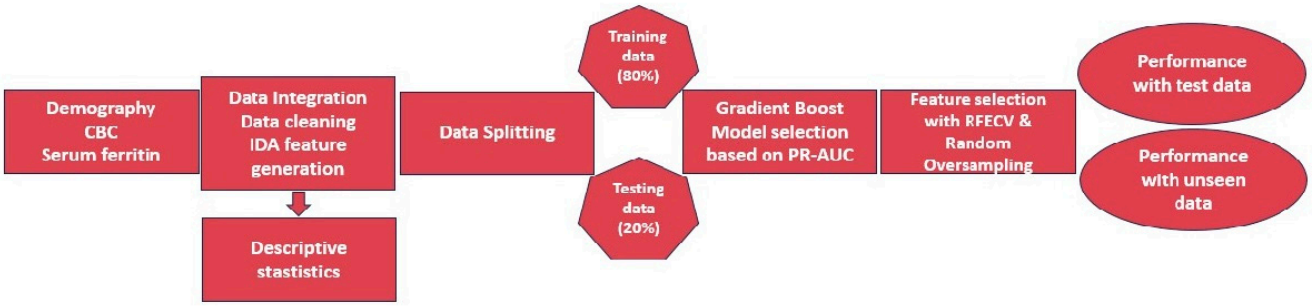

Figure 2. Workflow for the prediction of iron deficiency anemia using gradient boost machine learning model.

We then calculated the standard performance metrics for classification tasks, which are accuracy, precision, recall, and Area Under the Receiver Operating Characteristic Curve (ROC AUC). Accuracy measures the proportion of correct positive and correct negative predictions among all possible predictions. Recall, also known as sensitivity, or

the true positive rate, measures the proportion of accurate positive predictions among all possible positive predictions. Precision is the proportion of correct positive predictions. The formulas used are as follows.

$$\text{Accuracy}\left(y, \hat{y}\right) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1\left(\hat{y}_i = y_i\right)$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

Accuracy, Recall, and Precision ranged from 0 to 1. High Recall indicates that the model identifies most positive instances. Since there was a large class imbalance in the target class, accuracy alone is not considered a reliable indicator of model performance, so we also computed Precision Recall AUC (PR AUC) to choose an algorithm that best predicts both IDA and Non-IDA classes [37].

2.7. Feature Selection, Cross Validation, Model Explanation

Based on relative performance metrics, the Gradient Boost algorithm was chosen, and the best features were selected using recursive feature elimination (RFECV, scikit-learn library). We also tested the normalization of class imbalance via random oversampling from the scikit-learn library. This method randomly oversamples the minority class in the training dataset to balance the distribution of the classes. We then tested the trained model's generalizability by assessing its performance on unseen data from Kenya.

To provide explanations for the model, we calculated feature contribution to the prediction using the SHapley Additive exPlanations (SHAP) algorithm. SHAP feature values indicate the impact of each feature on the model's prediction for specific instances/classes. We also compared SHAP values of features with and without random over sampling to identify any differences to feature importance introduced by oversampling.

3. Results

3.1. Data Description, Proportion of Anemia, ID, and IDA

The sample distribution was 18.5%, 4.5%, 22.7%, and 54.2% for "under 5", 5–9, 10–19, and 19–60 age groups, respectively. Of the total sample, 76% were female subjects, and 5% of females were pregnant. The proportion (%) of anemia, ID, and IDA along with their 95% CIs by gender and age group are shown in Table 3 below. The overall proportion of anemia, ID and IDA are 8.6%, 14%, and 5%, respectively. The proportions of anemia (10.8% vs. 1.4%), ID (16.5% vs. 5.6%), and IDA (6.4% vs. 0.4%) are higher in females compared to males, and the proportions are higher in adolescents (10–19 years old) and adults (>19 years old) compared to young children aged less than 10 years old.

Table 3. Proportion of anemia, iron deficiency (ID), and iron deficiency anemia (IDA) by gender and age group.

Group	Anemia			ID			IDA		
	Proportion	**5% CI**	**95% CI**	**Proportion**	**5% CI**	**95% CI**	**Proportion**	**5% CI**	**95% CI**
All	8.63	8.24	9.02	13.99	13.51	14.47	4.99	4.69	5.29
Male	1.39	1.38	1.39	5.61	5.60	5.62	0.41	0.40	0.41
Female	10.86	10.85	10.86	16.57	15.56	16.57	6.40	6.40	6.41
<5 y	1.87	1.43	2.31	8.73	7.82	9.65	0.70	0.43	0.97
5–9 y	1.76	0.90	2.62	5.64	4.13	7.14	0.11	−0.10	0.32
10–19 y	9.23	8.39	10.07	16.15	15.08	17.21	5.13	4.49	5.77
>19 y	11.25	10.65	11.84	15.57	14.89	16.23	6.8	6.32	7.27

3.2. Comparative Analysis of Machine Learning Models

The classification of IDA and non-IDA in the dataset was tested using multiple machine learning models. The comparative prediction performance metrics are given in Table 4. The accuracy of all the models was 97% except Naïve Bayes which was 95%. The PR AUC was highest (0.87) using the Gradient boost model, which was therefore chosen for further optimization.

Table 4. Comparative prediction metrics of machine learning models on classification of anemia.

	Confusion Matrix	Accuracy	Precision	Recall	ROC AUC	PR AUC
Logistic regression	[[3769 22] [87 121]]	0.97	0.84	0.58	0.99	0.83
Random Forest	[[3761 30] [67 141]]	0.97	0.82	0.67	0.99	0.85
K-Nearest Neighbors	[[3759 32] [75 133]]	0.97	0.80	0.63	0.94	0.73
Naive Bayes	[[3656 135] [33 168]]	0.95	0.56	0.84	0.97	0.72
Gradient Boosting	[[3759 32] [61 147]]	0.97	0.82	0.70	0.99	0.87
XGBoost	[[3744 47] [55 153]]	0.97	0.76	0.73	0.99	0.85

3.3. Optimization of the Gradient Boost Algorithm and Cross Validation with Unseen Data

Feature selection was performed using the RFECV algorithm, with the selected features being LBXHGB, RIDAGEYR, LBXRDW, LBXMCHSI, RIAGENDR, RIDEXPRG, LBXHCT, and LBXMCVSI, and the performance of models with these features remained similar to that when all the features were included, except that a marginal increase in recall was observed (0.716 vs. 0.706). We used random oversampling to handle the class imbalance. In addition to the above features, LBDLYMNO and LBDMONO features were picked with RFECV; although this did not improve PR AUC significantly (0.87 vs. 0.87), there was a marked increase in recall/sensitivity (0.980 vs. 0.716). To assess how well the model performs in different settings, we evaluated its performance using a new dataset from Kenya. The accuracy of the model on this unseen dataset was found to be 0.98. Additionally, the precision of the model was 0.80, and the recall/sensitivity was 0.89 (Table 5).

Table 5. Gradient boost performance metrics with RFECV selected features, random oversampling, and validation on unseen data.

	Confusion Matrix	Accuracy	Precision	Recall	ROC AUC	PR AUC
With selected features	[[3761 30] [59 149]]	0.97	0.83	0.71	0.99	0.87
With selected features and random oversampling	[[3669 122] [4 204]]	0.96	0.62	0.98	0.99	0.87
Validation metrics with unseen data	[[479 4] [2 17]]	0.98	0.8	0.89	-	-

3.4. Influence of Oversampling on Model Explanations

The impact of the CBC features on the performance of the trained gradient boost model with and without random oversampling is given in Figure 3. In the SHAP summary plots (Figure 3 left panels), a positive SHAP value indicates an increase in the value of a specific feature associated with an increase in the model's prediction, while a negative relationship reduces the prediction of that instance, and the distance of the values from zero on the *x*-axis indicates the magnitude of its contribution to the prediction. A visualization using a red dot indicates a higher probability of indicating a positive class (IDA), while blue dots indicate the opposite (non-IDA). The explanatory model reveals that a low blood level of hemoglobin (LBXHGB), higher Age (RIDAGEYR), higher RDW (LBXRDW), being female and pregnant, and lower values of MCH, MCV, and HCT contribute to the prediction of IDA class, with their relative importance given (Figure 3, upper right panel). When trained with the random oversampling method, the feature importance was similar, except that the relative contribution of gender (RIAGENDR) was reduced, possibly due to a higher proportion of female subjects in the oversampled data due to a higher proportion of IDA class in this gender. The contribution of LBDLYMNO and LBDMONO was relatively lower compared to other features.

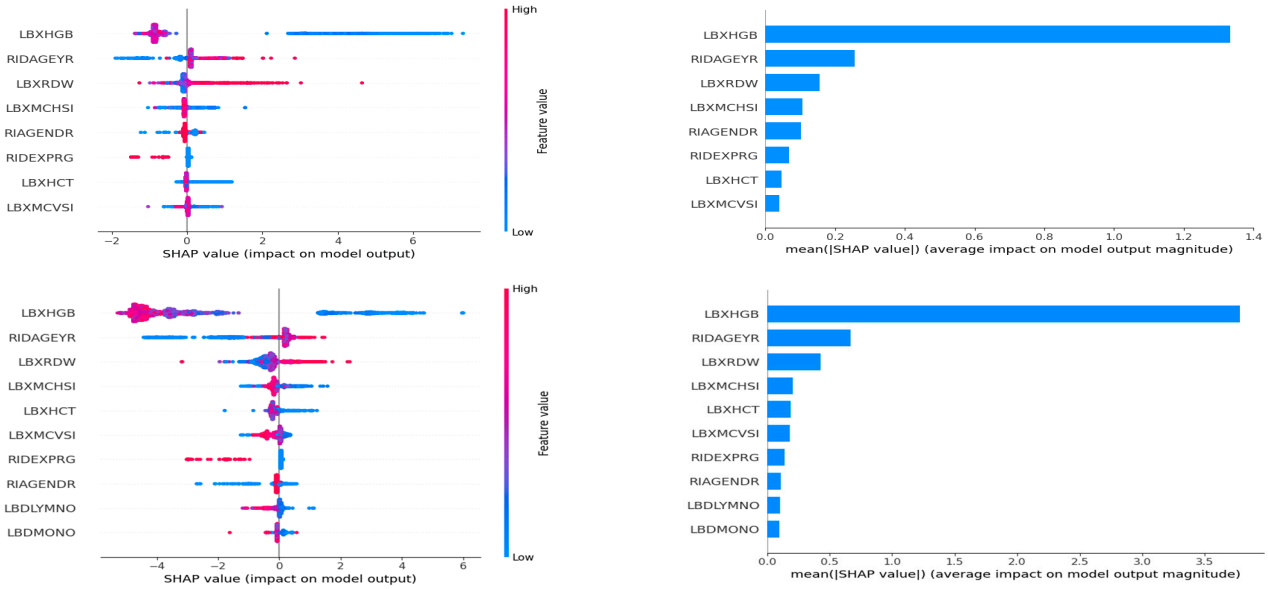

Figure 3. This figure shows SHAP summary plots (**left**) and feature importance (**right**) of model features and their average impact on the gradient boost IDA classification model without (**top**) and with (**bottom**) random oversampling.

4. Discussion

Since not all anemia is due to iron deficiency [2,4,38], the diagnosis of IDA normally requires both CBC (for hemoglobin) and serum ferritin tests [5]. Understanding the type of anemia is critical in making an informed decision on iron treatment either in hospital settings or in quantifying the proportion of IDA in survey settings. However, a serum ferritin test requires centrifugation of blood to collect serum or plasma and its transportation to the laboratory under a cold chain, and expensive lab infrastructure and expertise in handling are required to measure ferritin levels using enzyme-linked immunosorbent assays (ELISA) to identify IDA [5,33]. Since IDA has a unique effect on blood cell indices, such as reduced hemoglobin, HCT, and MCV and a higher RDW, it has been hoped that it would be possible to use this data for the precise identification of underlying iron deficiency without using the serum ferritin test [6–9], but there has been no large-scale study. In recent years, ML algorithms have found potential applications in medical diagnosis including classification based on numerical and image data [10], and several studies reported the

prediction of IDA based on CBC data (Supplementary Table S1). However, nearly all studies that specifically focused on IDA classification included serum iron or ferritin parameters in addition to CBC data for the prediction of IDA. One study predicted low ferritin (IDA) but only for data from adult anemic subjects (above 18 years old) generated from the referred lab tests. However, a model that applies to all age groups, including pregnant women, is lacking, possibly due to the extensive requirement of high-quality and paired CBC and serum ferritin data for the classification of IDA along with information on age, sex, and pregnancy status. Fortunately, these data are available in NHANES surveys, which provides an opportunity to test and classify IDA based on CBC variables.

In this study, we created a subset of data (n = 19,975) with paired demography, CBC, and serum ferritin values, and categorized the IDA based on their age and gender-specific hemoglobin and serum ferritin. In this dataset, we found that the proportion of subjects with IDA (5%) is close to half that with anemia (8.6%). Further, a higher proportion of anemia and IDA was in females compared to males, and the proportion of IDA among females also increased with age, which are consistent with reported age and gender differences in the prevalence of anemia and IDA [2].

Since there is severe class imbalance in the data, average precision (PR AUC) is considered the best performance metric in assessing model performance [37]. In our analysis, the Gradient Boost algorithm performed best and was chosen for the further optimization that included feature selection and class imbalance handling through random oversampling. Further, to test for differences if any, feature selection was performed without and with random oversampling using the RFECV algorithm, and the model performance was assessed with these selected features. The selected best features were the same with or without random oversampling, except that an additional variable was selected with the latter. However, random oversampling resulted in a marked increase in recall (sensitivity) for positive classes. Discussion with a domain expert suggested that it is important to have higher sensitivity for guiding treatment, and treating those without IDA or potentially borderline cases is unlikely to have large negative health impacts. The selected features such as age (RIDAGEYR), sex (RIAGENDR), and pregnancy status (RIDEXPRG) are expected as the reference values of hemoglobin and serum ferritin to classify anemia which is specific to these groups [1,5,33]. The other features related to hemoglobin and RBC morphology (LBXHGB, LBXHCT, LBXMCVSI, LBXMCHSI, LBXRDW) are also expected based on prior knowledge [9]. However, two additional WBC features, lymphocytes (LBDLYMNO) and monocytes (LBDMONO), which were not expected, also appear to contribute to the model's performance. Serum ferritin is acute phase protein, and inflammation elevates its levels independent of body iron stores [5,39]. Therefore, serum ferritin values are corrected for inflammation with the use of additional markers of inflammation such as C-reactive protein (CRP), prior to using it for assessing ID or IDA [39]. Since several WBC variables including lymphocyte and monocyte levels are elevated during inflammation [40], this explains their contribution to the model. Indeed, the lymphocyte count is (2.65 (95% CI 2.64–2.67) vs. 2.24 (95% CI 2.19–2.30)) and monocyte count is (0.57 (95% CI 0.56–574) vs. 0.55 (95% CI 0.54–0.56)). In addition, we also found that the importance of RIAGENDR (gender) was less when the model was trained with random oversampling. However, this difference could simply be due to oversampling of the data for this gender, with a higher proportion of positive IDA cases. Interestingly, when tested with unseen data originating from completely different settings, the trained model also performed exceptionally well, indicating its great utility across different datasets and/or geographics, and is stable to the subtle analytical differences across labs, if any.

Explaining disease diagnosis by showing the specific contributions of individual features to such prediction is of great importance to allow the clinician to make informed decisions. As explained above, the features selected in the trained algorithm are consistent with their known relationship with IDA. To further quantify the specific contribution of each feature to the prediction, we visualized the role of different features in the model using the SHAP algorithm. It was found that having a low blood level of hemoglobin (LBXHGB),

older age (RIDAGEYR), higher RDW (LBXRDW), being female and pregnant, and having lower values of MCH, MCV, HCT, lymphocytes, and monocytes contribute to the prediction of IDA class, with relative importances. While all these variable contributions are consistent with the known literature [6–8,27,28] the contribution of lower lymphocytes and monocytes to IDA could be mediated by the effect of inflammation on serum ferritin levels [5,39,41].

The published ML models for anemia classification all use different data than used here or perform a task other than discriminating IDA. For example, they rely on image data, or they use CBC variables to identify genetic disorders related to hemoglobin [10–20,27–30]. Since the cause of anemia is multifactorial [4], identifying concurrent iron deficiency is required to guide iron therapy. Our developed ML model is unique in identifying the presence of IDA based on CBC data alone, eliminating the need for an additional serum ferritin test. The diagnosis of IDA using our model has two important benefits. First, in a clinical setting, the model helps in making an informed decision to treat using iron supplementation. Second, in a public health survey setting, the model helps in quantifying the proportion of IDA, which contributes to understanding the factors associated with anemia, which is key information required for formulating appropriate actions.

5. Limitations and Future Directions

One limitation of our study is the serious class imbalance, which is typical of the disorder, although this imbalance can be overcome by using random over sampling, with minimal impact on the role of individual features. The other limitation is that we did not correct serum ferritin levels for inflammation prior to using it as a classifier of IDA, but we believe that this can partly be negated by CBC data on WBC indicators (lymphocytes and monocytes) which can be surrogate markers of inflammation. For example, WBC count has been shown to positively correlate with CRP levels [35], and future studies that compare the adjustment of serum ferritin using WBC indices with that of CRP corrected values merits investigation.

We also acknowledge that testing the model's performance in a real clinical setting would be beneficial, but this is beyond the scope of the study. However, performing a clinical trial could be a logistical extension of this ML model, where it might be directly integrated into the CBC analyzer itself with the results validated via serum ferritin tests (reverse of what is performed now). Since conducting such a trial would be costly, evidence such as this data study would likely be needed to justify it, and this testing would most likely be carried out by blood analysis device manufacturers.

6. Conclusions

Our study successfully developed a machine learning model that discriminates IDA from other forms of anemia, solely based on CBC data with particularly high recall for IDA positive cases. The study used two publicly available datasets (a NHANES dataset of more than 19,000 instances from the US and a smaller dataset from Kenya) and found that IDA can be classified from CBC data with a PR AUC of 0.87 and recall/sensitivity of 0.98 and 0.89 for the original dataset and the unseen one collected in Kenya, respectively. In the analysis of feature importance, we found that a low blood level of hemoglobin (LBXHGB), older age (RIDAGEYR), higher RDW (LBXRDW), being female and pregnant, and lower values of MCH, MCV, HCT, lymphocytes, and monocytes contributed most to the prediction of IDA class. The precision, recall, and generalizability of the model to unseen data and the role of features associated with the predictions align well with known prior knowledge, which strongly supports the confidence in providers making informed decisions on whether to treat with iron or advise additional testing. It would also be useful in public health survey settings to quantify the proportion of iron deficiency associated with anemia, without additional testing. In the future, following appropriate clinical trials, it would be possible to integrate this model directly into CBC analyzer platforms to provide diagnostic decision support along with CBC reporting, which should encourage the development of enhanced CBC platforms.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/biomedinformatics4010036/s1, Table S1 Comparison of reported machine learning models for prediction of iron deficiency anemia.

Author Contributions: Conceptualization, S.P. and S.M.; data curation, S.P.; formal analysis, S.P.; investigation, S.P.; methodology, S.P. and S.M.; project administration, S.M.; resources, S.P. and S.M.; software, S.P.; supervision, S.M.; validation, S.P. and S.M.; visualization, S.P.; writing—original draft, S.P. and S.M.; writing—review & editing, S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data from this study are derived from the NAHNES dataset. The original dataset can be found at (https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx, accessed on 21 February 2024). For reproducibility our code and cleaned data are online and public. Access our repository on GitHub at: https://github.com/siddartha-10/Classification_of_IDA.

Acknowledgments: The authors acknowledge the contributions of Ravindranadh Palika Scientist-C, ICMR-National Institute of Nutrition, Hyderabad, India for domain-specific suggestions and inputs to the interpretation of the findings. The authors also acknowledge the support of the Office of the Dean of the College of Engineering & Applied Science at the University of Wisconsin-Milwaukee.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. World Health Organization (WHO). *Haemoglobin Concentrations for the Diagnosis of Anaemia and Assessment of Severity*; WHO: Geneva, Switzerland, 2011.
2. GBD 2021 Anaemia Collaborators. Prevalence, Years Lived with Disability, and Trends in Anaemia Burden by Severity and Cause, 1990–2021: Findings from the Global Burden of Disease Study. *Lancet Haematol.* **2023**, *10*, e713–e734. [CrossRef] [PubMed]
3. Hsia, C.C. Respiratory Function of Hemoglobin. *N. Engl. J. Med.* **1998**, *338*, 239–247. [CrossRef] [PubMed]
4. Sarna, A.; Porwal, A.; Ramesh, S.; Agrawal, P.K.; Acharya, R.; Johnston, R.; Khan, N.; Sachdev, H.P.S.; Nair, K.M.; Ramakrishnan, L.; et al. Characterisation of the Types of Anaemia Prevalent among Children and Adolescents Aged 1-19 Years in India: A Population-Based Study. *Lancet Child Adolesc. Health* **2020**, *4*, 515–525. [CrossRef] [PubMed]
5. Zimmermann, M.B.; Hurrell, R.F. Nutritional Iron Deficiency. *Lancet* **2007**, *370*, 511–520. [CrossRef] [PubMed]
6. Uchida, T. Change in Red Blood Cell Distribution Width with Iron Deficiency. *Clin. Lab. Haematol.* **1989**, *11*, 117–121. [CrossRef] [PubMed]
7. van Zeben, D.; Bieger, R.; van Wermeskerken, R.K.A.; Castel, A.; Hermans, J. Evaluation of Microcytosis Using Serum Ferritin and Red Blood Cell Distribution Width. *Eur. J. Haematol.* **1990**, *44*, 106–109. [CrossRef] [PubMed]
8. Burk, M.; Arenz, J.A.; Schneider, W. Erythrocyte Indices as Screening Tests for the Differentiation of Microcytic Anemias. *Eur. J. Med. Res.* **1995**, *1*, 33–37.
9. Cascio, M.J.; DeLoughery, T.G. Anemia: Evaluation and Diagnostic Tests. *Med. Clin.* **2017**, *101*, 263–284. [CrossRef]
10. Kang, M. Machine Learning: Diagnostics and Prognostics. *Progn. Health Manag. Electron.* **2018**, 163–191. [CrossRef]
11. Al-Zaiti, S.; Martin-Gill, C.; Zègre-Hemsey, J.; Medicine, Z.B. Machine Learning for ECG Diagnosis and Risk Stratification of Occlusion Myocardial Infarction. *Nat. Med.* **2023**, *29*, 1804–1813. [CrossRef]
12. Ayyıldız, H.; Tuncer, S.A. Determination of the Effect of Red Blood Cell Parameters in the Discrimination of Iron Deficiency Anemia and Beta Thalassemia via Neighborhood Component Analysis. *Chemom. Intell. Lab. Syst.* **2020**, *196*, 103886. [CrossRef]
13. Vohra, R.; Hussain, A.; Dudyala, A.K.; Pahareeya, J.; Khan, W. Multi-Class Classification Algorithms for the Diagnosis of Anemia in an Outpatient Clinical Setting. *PLoS ONE* **2022**, *17*, e0269685. [CrossRef] [PubMed]
14. Khan, J.R.; Chowdhury, S.; Islam, H.; Raheem, E. Machine Learning Algorithms to Predict the Childhood Anemia in Bangladesh. *J. Data Sci.* **2019**, *1*, 195–218. [CrossRef]
15. Dejene, B.E.; Abuhay, T.M.; Bogale, D.S. Predicting the Level of Anemia among Ethiopian Pregnant Women Using Homogeneous Ensemble Machine Learning Algorithm. *BMC Med. Inform. Decis. Mak.* **2022**, *22*, 247. [CrossRef] [PubMed]
16. Appiahene, P.; Asare, J.W.; Donkoh, E.T.; Dimauro, G.; Maglietta, R. Detection of Iron Deficiency Anemia by Medical Images: A Comparative Study of Machine Learning Algorithms. *BioData Min.* **2023**, *16*, 2. [CrossRef]
17. Jain, P.; Bauskar, S.; Gyanchandani, M. Neural Network Based Non-Invasive Method to Detect Anemia from Images of Eye Conjunctiva. *Int. J. Imaging Syst. Technol.* **2020**, *30*, 112–125. [CrossRef]

18. Jayakody, J.A.; Edirisinghe, E.A. HemoSmart: A Non-Invasive, Machine Learning Based Device and Mobile App for Anemia Detection. In Proceedings of the 2020 IEEE Region 10 Conference (TENCON), Osaka, Japan, 16–19 November 2020; pp. 1401–1406. [CrossRef]
19. Asare, J.W.; Appiahene, P.; Donkoh, E.T.; Dimauro, G. Iron Deficiency Anemia Detection Using Machine Learning Models: A Comparative Study of Fingernails, Palm and Conjunctiva of the Eye Images. *Eng. Rep.* **2023**, *5*, e12667. [CrossRef]
20. Sen, B.; Ganesh, A.; Bhan, A.; Dixit, S.; Goyal, A. Machine Learning Based Diagnosis and Classification of Sickle Cell Anemia in Human RBC. In Proceedings of the 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 4–6 February 2021; pp. 753–758. [CrossRef]
21. Bellinger, C.; Amid, A.; Japkowicz, N.; Victor, H. Multi-Label Classification of Anemia Patients. In Proceedings of the 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 9–11 December 2015; pp. 825–830. [CrossRef]
22. Saputra, D.C.E.; Sunat, K.; Ratnaningsih, T. A New Artificial Intelligence Approach Using Extreme Learning Machine as the Potentially Effective Model to Predict and Analyze the Diagnosis of Anemia. *Healthcare* **2023**, *11*, 697. [CrossRef]
23. Dogan, S.; Turkoglu, I. Iron-Deficiency Anemia Detection from Hematology Parameters by Using Decision Trees. *Int. J. Sci. Technol.* **2008**, *3*, 85–92.
24. Azarkhish, I.; Raoufy, M.R.; Gharibzadeh, S. Artificial Intelligence Models for Predicting Iron Deficiency Anemia and Iron Serum Level Based on Accessible Laboratory Data. *J. Med. Syst.* **2012**, *36*, 2057–2061. [CrossRef]
25. Yilmaz, A.; Dagli, M.; Allahverdi, N. A Fuzzy Expert System Design for Iron Deficiency Anemia. In Proceedings of the 2013 7th International Conference on Application of Information and Communication Technologies, Baku, Azerbaijan, 23–25 October 2013; pp. 1–4. [CrossRef]
26. Yıldız, T.K.; Yurtay, N.; Öneç, B. Classifying Anemia Types Using Artificial Learning Methods. *Eng. Sci. Technol. Int. J.* **2021**, *24*, 50–70. [CrossRef]
27. Terzi, E.; Sarıbacak, B.; Sağlam, F.; Cengiz, M.A. A Novel Expert System for Diagnosis of Iron Deficiency Anemia. *Comput. Math Methods Med.* **2022**, *2022*, 7352096. [CrossRef] [PubMed]
28. Kurstjens, S.; De Bel, T.; Van Der Horst, A.; Kusters, R.; Krabbe, J.; Van Balveren, J. Automated Prediction of Low Ferritin Concentrations Using a Machine Learning Algorithm. *Clin. Chem. Lab. Med.* **2022**, *60*, 1921–1928. [CrossRef] [PubMed]
29. Nashwan, A.J.; Alkhawaldeh, I.M.; Shaheen, N.; Albalkhi, I.; Serag, I.; Sarhan, K.; Abujaber, A.A.; Abd-Alrazaq, A.; Yassin, M.A. Using Artificial Intelligence to Improve Body Iron Quantification: A Scoping Review. *Blood Rev.* **2023**, *62*, 101133. [CrossRef] [PubMed]
30. Yang, C.C. Explainable Artificial Intelligence for Predictive Modeling in Healthcare. *J. Healthc. Inform Res.* **2022**, *6*, 228–239. [CrossRef] [PubMed]
31. NHANES—National Health and Nutrition Examination Survey Homepage. Available online: https://www.cdc.gov/nchs/nhanes/index.htm (accessed on 12 February 2024).
32. Pandas Documentation—Pandas 2.2.0 Documentation. Available online: https://pandas.pydata.org/docs/ (accessed on 12 February 2024).
33. World Health Organization (WHO). *WHO Guideline on Use of Ferritin Concentrations to Assess Iron Status in Populations*; WHO: Geneva, Switzerland, 2020.
34. Patel, K. V Epidemiology of Anemia in Older Adults. *Semin. Hematol.* **2008**, *45*, 210–217. [CrossRef] [PubMed]
35. Omuse, G.; Chege, A.; Kawalya, D.E.; Kagotho, E.; Maina, D. Ferritin and Its Association with Anaemia in a Healthy Adult Population in Kenya. *PLoS ONE* **2022**, *17*, e0275098. [CrossRef] [PubMed]
36. Omuse, G.; Maina, D.; Mwangi, J.; Wambua, C.; Radia, K.; Kanyua, A.; Kagotho, E.; Hoffman, M.; Ojwang, P.; Premji, Z.; et al. Complete Blood Count Reference Intervals from a Healthy Adult Urban Population in Kenya. *PLoS ONE* **2018**, *13*, e0198444. [CrossRef]
37. Saito, T.; Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef]
38. Pasricha, S.R.; Armitage, A.E.; Prentice, A.M.; Drakesmith, H. Reducing Anaemia in Low Income Countries: Control of Infection Is Essential. *BMJ* **2018**, *362*. [CrossRef]
39. Namaste, S.M.; Rohner, F.; Huang, J.; Bhushan, N.L.; Flores-Ayala, R.; Kupka, R.; Mei, Z.; Rawat, R.; Williams, A.M.; Raiten, D.J.; et al. Adjusting Ferritin Concentrations for Inflammation: Biomarkers Reflecting Inflammation and Nutritional Determinants of Anemia (BRINDA) Project. *Am. J. Clin. Nutr.* **2017**, *106* (Suppl. S1), 359S–371S. [CrossRef]
40. Oda, E.; Kawai, R. Comparison between High-Sensitivity C-Reactive Protein (Hs-CRP) and White Blood Cell Count (WBC) as an Inflammatory Component of Metabolic Syndrome in Japanese. *Intern. Med.* **2010**, *49*, 117–124. [CrossRef]
41. Seo, I.H.; Lee, Y.J. Usefulness of Complete Blood Count (CBC) to Assess Cardiovascular and Metabolic Diseases in Clinical Settings: A Comprehensive Literature Review. *Biomedicines* **2022**, *10*, 2697. [CrossRef]