

## Retrospective Examination of Risk Factors Affecting Iron Deficiency Anemia Using Machine Learning Methods

Erol Terzi <sup>1,a,\*</sup>, Bünyamin Sarıbacak <sup>2,b</sup>, Mehmet Şirin Ateş <sup>1,c</sup>

<sup>1</sup> Department of Statistics, Faculty of Science, Ondokuz Mayıs University, Samsun, Türkiye.

<sup>2</sup> Department of Computer Science, Education Faculty, Ondokuz Mayıs University, Samsun, Türkiye.

\*Corresponding author

### Research Article

#### History

Received: 13/07/2023

Accepted: 22/06/2024





This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)


### ABSTRACT


Iron deficiency anemia is one of the most common types of anemia worldwide. In recent years, new developments in the field of medicine have offered early diagnosis and treatment opportunities for anemia patients. In the field of data science, in parallel with the developments in medicine, significant developments are taking place in subjects such as data collection, storage, processing, and reporting. Interdisciplinary joint studies positively contribute to patients' quality of life and lifespan. In this study, the accuracy of the statistical results was tested with Machine Learning Method (MLM) while investigating the factors that affect the correct prediction of Iron Deficiency Anemia (IDA) diagnosis. In the first stage, the relationships between all variables in the data set and their effects on the differentiation of disease groups were investigated using univariate and multivariate statistical methods. In the second step, the data set was analyzed in detail using four different methods with Artificial Neural Network (ANN) classifier. Weka 3.8 application was preferred for these operations. In the last stage, the results obtained in both stages were compared. Accordingly, it has been observed that hemoglobin (Hb), mean cell volume (MCV), iron (Fe), and ferritin (FERR) have more effects on IDA. ANN (98.06%) is a better discriminator with a correct classification rate.


**Keywords:** Iron deficiency anemia, Logistic regression analysis, Artificial neural network, Machine learning.

 [eroltr@omu.edu.tr](mailto:eroltr@omu.edu.tr)

 [mehmet.ates@omu.edu.tr](mailto:mehmet.ates@omu.edu.tr)

 <https://orcid.org/0000-0002-2309-827X>

 <https://orcid.org/0000-0001-9904-6380>

 [bunyamin@omu.edu.tr](mailto:bunyamin@omu.edu.tr)

 <https://orcid.org/0000-0003-2775-776X>

## Introduction

Iron deficiency is an essential determinant of anemia and is among the most common types worldwide, according to the World Health Organization (WHO) [1]. Any form of malnutrition has been shown as the main cause of anemia. Despite all efforts to combat malnutrition worldwide, progress has been limited. 614 million women and 280 million children worldwide are struggling with this disease. In particular, it affects 40% of pregnant women, 33% of non-pregnant women, and 42% of children worldwide [1]. In addition to significant and irreversible effects on brain development in children under two, iron deficiency negatively affects learning and school performance in later life [2,3]. Iron deficiency in adults has negative effects such as fatigue, physical performance impairment, and decreased work efficiency and social activities [4,5].

The global prevalence of anemia for the general population is 24.8%, and it is estimated that 1 billion 620 million people are affected by anemia [1]. Despite the positive developments in healthcare services worldwide, the number of people affected by anemia has increased, primarily due to population growth. This result may be due to the ineffective evaluation of the blood analysis given by the clinician whose primary interest is outside this field or the ineffective treatment of anemia.

This study aims to estimate the risk factors affecting Iron Deficiency Anemia (IDA) with acceptable sensitivity. Today, technological developments in the field of computers and especially artificial intelligence are used extensively to solve problems in the field of health [6,7]. Machine learning, a sub-branch of artificial intelligence, consists of systematic techniques developed to make accurate predictions using previous observations [8]. This study establishes a harmonious relationship among statistics, databases and machine learning disciplines. In the first step, univariate statistical analyses were carried out to investigate the effects of selected parameters on individual disease groups. In the second step, logistic regression multivariate statistical analysis was performed in which all parameters found to be significant were added to the model to predict the probability of an outcome with only two values. In the last step, the data were analyzed using the Artificial Neural Network (ANN) machine learning classifier, which includes mathematical methods for analyzing nonlinear functions [9]. It has been investigated which model is more successful in diagnosing the consequences of risk factors affecting IDA.

The paper proceeds as follows. Section 2 describes the Logistic Regression (LR) and ANN models. The third section presents result from the models estimation. A more detailed evaluation of the results is made in the last

section. It is anticipated that the results obtained will be a reference for future scientific studies in this field.

**Materials and Method**

**Collecting Data**

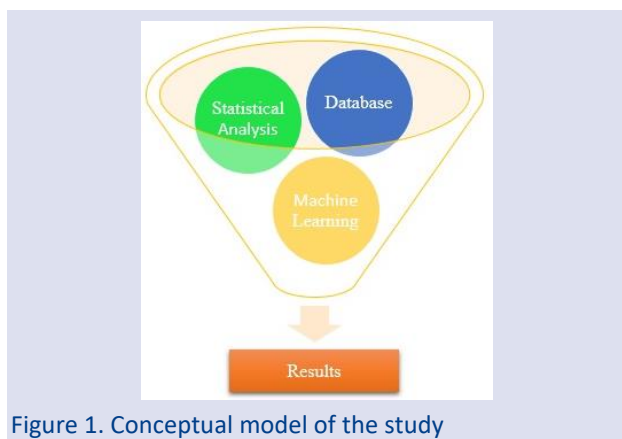
Between October 2017 and March 2020, 516 patients diagnosed with malaise and fatigue (ICD-10 code: R53) who applied to the Samsun Training and Research Hospital Hematology Outpatient Clinic were analyzed retrospectively. IDA was diagnosed in 359 patients according to laboratory results. Although the remaining 157 cases were evaluated with the same diagnosis, laboratory values did not give same result as IDA.

Age, gender, complete blood count (CBC) values, Hb, hematocrit (Hct), mean cell volume (MCV), mean cell Hb concentration (MCHC), red cell distribution width (RDW), red blood cell (RBC) and IDA parameters (ferritin, serum iron, serum iron binding capacity, transferrin saturation) of these 516 cases were recorded. The abbreviations above are shown in Table 1.

Table 1. List of laboratory test abbreviations used in this study

Laboratory Test	Abbreviations
Hemoglobin	Hb
Hematocrit	Hct
Mean Cell Volume	MCV
Mean Corpuscular Hemoglobin Concentration	MCHC
Red Blood Cell Count	RBC
Red Blood Cell Distribution Width	RDW
Iron	Fe
Unsaturated Iron Binding Capacity	UIBC
Ferritin	FERR
Disease Diagnosis	DD

Statistical analysis and machine learning techniques were applied to the patients’ data to verify these analyses. The conceptual model of the study is shown in Figure 1.



**Statistical Analysis**

In the statistical analysis of this study, SPSS 21.0 software was preferred. In all statistical analyzes, the level of significance was accepted as  $\alpha = 0.05$ . Frequency tables and descriptive statistics were used in the interpretation of the findings. When the data were

normally distributed was used “Independent Sample t test” which is one of the parametric tests and when the data were not normally distributed was used “Mann Whitney U test” which is one of the non-parametric tests. In addition, “Binary Logistic Regression (LR)” model was used to classification the factors affecting the disease risk Logistic regression predicts the probability of an outcome that can only have two values [10].

$$Odds = \frac{p}{1 - p} \tag{1}$$

The Odds ratio is the ratio of the Odds coefficients of two cases investigated. This ratio is used to explain the effect of the dependent variable on independent variables [11,12]. The basic concept in this model is Logit, and this is shown as

$$Logit(p) = \ln\left(\frac{p}{1-p}\right) \tag{2}$$

Logit(p) takes negative values when the probability takes values below 0.5, and Logit(p) takes positive values when the probability takes values above 0.5. The higher the probability value, the higher the Logit(p) value [6]. Odds ratio has been widely used in recent years due to its ability to give outstanding estimates with the help of confidence intervals for relationships between binary variables, to investigate the effects of independent variables in studies using logistic regression, and to achieve successful results in case-control studies [13].

**Artificial Neural Network (ANN)**

ANN is an information-processing technique inspired by the way the human brain works. ANN is widely used today in many applications in various branches of engineering, medicine, and science [14]. ANN consists of connecting artificial neurons called nodes. These connections are assigned a value according to their strength, and the higher the value, the stronger the connection. ANN consists of input, hidden, and exit nodes (Figure 2). ANN networks have forward and feedback features. In the feed-forward, signals in ANN move only in one direction, and the values calculated from the input data are the input values of the next layer. This process continues in all layers until output. They can move in both directions by using loop structures in feedback ANN. They try to create the most appropriate connection by controlling all possible connections between neurons. Using a backpropagation learning algorithm, this study used the hematological dataset to train a Multilayer Perceptron (MLP).

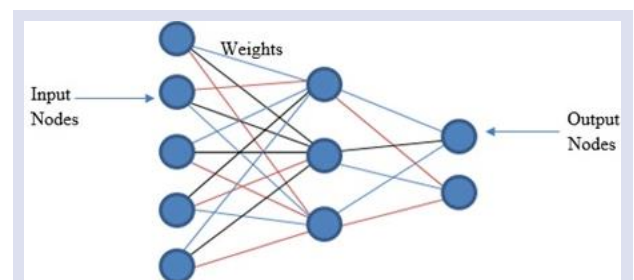


Figure 2. Artificial neural network [17]

## Results

In this study, in which risk factors affecting iron deficiency anemia were examined, information on 516 patients subject to the search is given in Table 2. Beforehand, the normality assumption of the data was

analyzed using the Kolmogorov Smirnov test. Data are normally distributed. (n>30)

Table 2. Comparison of some parameters according to the groups

Variable (N=516)	Healthy (n=157)		Patient (n=359)		Statistical Analysis* Probability
	$\bar{X} \pm SD$	Median [Min-Max]	$\bar{X} \pm SD$	Median [Min-Max]	
Age (year)	54,37±14,40	54,0 [21,0-89,0]	43,29±14,11	41,0 [17,0-87,0]	Z=-7,969 p=0,000
Hb	13,38±1,28	13,2 [11,0-17,0]	10,32±1,79	10,3 [6,2-15,8]	t=21,906 p=0,000
Hct	38,84±3,62	38,5 [32,0-48,8]	31,20±4,76	31,1 [18,3-48,2]	t=19,955 p=0,000
MCV	84,90±6,38	85,2 [34,9-99,7]	71,76±9,34	71,9 [51,9-111,6]	Z=-14,272 p=0,000
MCH	29,45±2,27	29,4 [21,5-38,3]	23,75±3,86	23,9 [14,6-40,1]	Z=-14,500 p=0,000
MCHC	34,32±1,07	34,3 [25,1-36,2]	32,99±1,34	33,1 [28,2-36,2]	Z=-11,158 p=0,000
RBC	4,50±0,53	4,5 [3,1-5,9]	4,37±0,58	4,4 [2,0-6,6]	Z=-2,650 p=0,008
RDW	15,21±3,48	14,1 [12,0-31,5]	18,39±4,43	17,3 [12,1-46,9]	Z=-11,167 p=0,000
FE	86,49±26,81	80,0 [40,0-217,0]	32,36±38,64	23,0 [10,0-464,0]	Z=-16,174 p=0,000
UIBC	252,68±50,93	251,0 [135,0-404,0]	390,35±96,81	405,0 [68,0-717,0]	Z=-13,987 p=0,000
Saturation (%)	0,36±0,16	0,3 [0,1-0,9]	0,12±0,31	0,1 [0,0-4,6]	Z=-16,150 p=0,0000
FERR	108,68±79,95	83,0 [28,0-640,0]	52,59±170,94	7,0 [10,0-1650,0]	Z=-14,382 p=0,000

\* “Independent Sample-t” test (t-table value) in comparison with the measurement values of two independent groups with normal distribution; “Mann-Whitney U” test (Z-table value) statistics were used in comparing the measurement values of two independent groups with no normal distribution

Table 3. Analysis of the relationships between groups and gender

Variable (N=516)	Healthy (n=157)		Patient (n=359)		Statistical Analysis* Probability
	n	%	n	%	
Female	107	68,2	332	92,5	$\chi^2=49,016$ p=0,000
Male	50	31,8	27	7,5	

\* The  $\chi^2$ -cross table investigated the relationships between two qualitative variables.

A statistically significant relationship was found between the groups and gender ( $\chi^2=49.016$ ;  $p=0.000$ ). It was determined that 50 people (31.8%) in the healthy

group were men, and 332 (92.5%) in the patient group were women.

The created model has a 93.8% correct classification rate (CCR). As a result of Logistic regression (Backward: LR method) performed to determine the disease risk status by including all parameters that were significant in the univariate analysis according to the groups (Table 2), the optimal model consists of the age (year), Hb, MCV, erythrocyte RBC, RDW, Fe, UIBC and saturation (%) parameters. Age, Hb, MCV, Erythrocyte RBC, RDW, Fe, and UIBC values were significant in the model ( $p < 0.05$ ). Some analysis results of the ANN classifier selected to compare the results of univariate and multivariate statistical analyzes are given in Table 5.

Table 4. Examination of the factors affecting the disease status with the LR model.

Variable	β	Standard Error	Wald	df	p	OR	95% Confidence Interval (CI)	
							Lower	Upper
							Age (year)	-0,057
Hb	-2,710	0,587	21,305	1	0,000	0,067	0,021	0,210
MCV	0,218	0,099	4,832	1	0,028	1,243	1,024	1,509
RBC	4,614	1,522	9,196	1	0,002	1,091	1,011	1,991
RDW	0,123	0,059	4,360	1	0,037	1,131	1,008	1,269
Fe	-0,044	0,016	7,129	1	0,008	0,957	0,927	0,988
UIBC	0,022	0,005	20,919	1	0,000	1,022	1,013	1,032
Saturation (%)	5,726	3,217	3,167	1	0,075	3,068	0,560	16,808
Constant	-8,316	8,537	0,949	1	0,330	0,000		

\*Hosmer & Lemeshow test  $\chi^2=2.236$ ;  $p=0.973$ ; CCR=93.8%

Table 5. Comparative analysis of ANN classifier

Classification output	Model 1	Model 2	Model 3	Model 4
Correctly Classified Instances	93.15%	96.9%	93.03%	98.06%
In Correctly Classified Instances	6.85%	3.1%	6.97%	1.94%
Kappa statistic	0.83	0.928	0.8358	0.955
Mean absolute error (MAE)	0.0802	0.048	0.0808	0.0405
Root mean squared error(RMSE)	0.2466	0.1633	0.2363	0.1451
Relative absolute error (RAE)	19.14%	11.34%	19.07%	9.57%
Root relative squared error	54.98%	35.5%	51.37%	31.54%
AUC	0.94	0.97	0.93	0.98

In Model-1, before the classification process, the data set is divided into 66% training set and 34% test set. In Model-2, the entire data set was used as a training set. In Model 3, the dimension was reduced with Principal Component Analysis (PCA) to highlight the strengths of the data before classification, and then the data set was divided into ten parts by cross-validation method, and nine pieces of the training set were used as one piece of the test set. Average accuracy was obtained as a result of 10 training and testing processes. In Model 4, first, a feature subset (Hb, MCV, Fe, and FERR) was determined by considering the individual prediction ability of each feature and the degree of excess between them. Afterward, a new classification was made in which only the specified features were included in the model, and the entire data set was used as a training set [15]. At the end of this process, the highest correct classification rate was obtained at 98.06%. The evaluation module also provides the correct classification rates as well as Kappa statistics, MAE, RMSE, and relative error (RAE) based on prior probabilities, and it also gives the statistics of the root mean square (RMSE) of the mean square (quadratic) loss [16].

While determining the disease groups in Tables 6 and 7, the ANN classifier uses the Sigmoid function as the transfer function. Nodes consist of the sum of the weights of their inputs. This function was preferred because the value ranges of logistic regression and sigmoid function are [0-1].

Table 6. Power of features to determine disease groups according to ANN

Inputs	Weights						
	Attributes	Node 2	Node 3	Node 4	Node 5	Node 6	Node 7
Threshold	-3,0879	-1,3453	-0,2288	-4,3075	0,3793	-8,4297	
age	-3,3420	-4,5974	0,1956	-5,2892	2,3902	-0,0890	
Hb	-6,1126	-5,2270	0,1856	-2,4230	-10,5547	-3,1134	
Hct	-1,1477	-1,5082	0,1452	-0,4477	-8,5774	-0,3969	
MCV	2,2419	1,6541	-0,6381	4,5104	-0,4725	6,7628	
MCH	-3,5715	-3,1097	-0,2216	-2,3489	-3,7943	-2,9190	
MCHC	2,6999	1,0754	0,2470	4,1382	-0,5154	8,2017	
RBC	3,8531	2,7593	-0,1178	2,1436	0,2101	3,4655	
RDW	1,9810	1,9663	0,5450	1,4202	0,5456	-2,1367	
FE	6,7096	5,7775	-0,1225	-4,8643	-0,4588	-23,4869	
UIBC	6,0031	3,1920	0,6698	-4,4654	-2,7588	3,1527	
FERR	-12,9757	-6,8534	0,1784	-6,1465	5,1607	5,7663	

Considering the individual prediction ability of each feature and the degree of excess between them, it is seen that the power of Hb, MCV, Fe, and FERR to determine disease groups is higher.

Table 7. The effects of nodes on disease groups according to ANN

Inputs	Node 0 - Class DD(Yes)	Node 1 - Class DD (No)
	Weights	
Threshold	-7,1850	7,1827
Node 2	6,7132	-6,7180
Node 3	4,6409	-4,6325
Node 4	-0,7723	0,7866
Node 5	4,7579	-4,7584
Node 6	8,3808	-8,3813
Node 7	11,3668	-11,3680

It can be observed that Nodes 2, 6, and 7 have a higher influence on the formation of disease classes.

## Conclusion

In this study, the power of age, gender, and results obtained from complete blood count (CBC) analysis in evaluating the diagnosis of IDA of individuals who applied to the hematology service was investigated. RBC indexes consisting of RBC, Hb, Hct, MCV, MCH, MCHC, Fe, UIBC, FERR, and RDW are produced from CBC analysis using automatic devices. According to the first information we obtained from the CBC results, it was found that the age (years), Hb, Hct, MCV, MCH, MCHC, Erythrocyte RBC, Fe, saturation (%), and FERR values of the patients diagnosed with IDA were lower than the healthy group. Likewise, the RDW and UIBC values of the patient group were statistically significantly higher than the healthy group, which is expected (Table 2). It was determined that the women subject to the study were predominantly ill, while the men were predominantly healthy (Table 3). According to the results of Logistic regression performed by including all parameters that were significant according to the groups as a result of univariate analyses (Table 2) to determine the disease risk status, when age (year) increases by 1 unit, the risk of becoming sick 5.6% will decrease by 5.6. When the Hb and MCV value increase by 1 unit, the risk of getting sick will reduce by 93.3% and 24.3%, respectively. When the erythrocyte RBC value increases by 0.01 units, the risk of becoming sick will increase by 9.1%. When the RDW, Iron, and UIBC values increase by 1 unit, the risk of getting sick will increase by 13.1%, 4.4%, and 2.2%, respectively (Table 4). The results of both univariate (Table 2) and multivariate (Table 4) analyses showed that only one parameter was insufficient to distinguish between two conditions; on the contrary, all parameters effectively separated the two groups. Hb, MCV, RBC, RDW, and Fe were found to be stronger separators in multiple comparison tests (Table 4).

The accuracy of these results was tested by classifying it with machine learning ANN using four methods (Table 5). As a result, it was observed that the correct classification rates of the new subsets, which were compressed and formed by reducing the number of features, changed positively. Although there is a low correlation between the elements of these clusters, it has been observed that there is a high degree of correlation in the classification. As a result of these analyses, the best correct classification rates of LR (93.8%) and ANN (98.06%) were determined, and it was seen that the ANN classifier was a better discriminator in determining the disease groups. The results highly confirm our predictions. It is hoped that advances in data science will significantly contribute to eliminating the difficulty of separating the factors affecting IDA. It is thought that new machine learning techniques to be applied to data sets containing large-scale current IDA disease data free of regional differences will better determine the risk factors affecting anemia.

For this, researchers (medical doctors, data scientists, etc.) who conduct interdisciplinary collaborations will

have detailed knowledge about diseases during the data collection, contributing significantly to the data analysis stages. Reducing the number of variables defined initially will greatly facilitate physicians' decision-making processes. In addition, converting the results obtained through many stages into easy-to-use digital applications will significantly reduce the loss of time and effort.

## Conflict of Interest

There are no conflicts of interest in this work.

## References

- [1] World Health Organization. *Anemia*, (2017), [https://www.who.int/health-topics/anaemia#tab=tab\\_1](https://www.who.int/health-topics/anaemia#tab=tab_1), Accessed 18 Nov 2022.
- [2] Allali S., Brousse V., Sacri A. S., Chalumeau M., De Montalembert M., Anemia in Children: Prevalence, Causes, Diagnostic Work-up, and Long-Term Consequences, *Expert Review of Hematology*, 10(11) (2017) 1023-1028.
- [3] Cusick S. E., Georgieff M. K., Rao R., *Approaches for Reducing the Risk of Early-Life Iron Deficiency-Induced Brain Dysfunction in Children*, *Nutrients* 10 (2) (2018) 227.
- [4] Andro M., Le Squire P., Estivin S., Gentric A., Anaemia and Cognitive Performances in The Elderly: A Systematic Review, *European Journal of Neurology*, 20(9) (2013) 1234-1240.
- [5] Haas J.D., Brownlie T., 4th. Iron Deficiency and Reduced Work Capacity: A Critical Review of the Research to Determine a Causal Relationship, *Journal of Nutrition*, 131 (2S-2) (2001) 676-690.
- [6] Hosmer D.W., Lemeshow S., *Applied Logistic Regression*, John Wiley & Sons, 8-36, New York, 1989.
- [7] Khan J. R., Chowdhury S., Islam H., Raheem E., Machine Learning Algorithms to Predict the Childhood Anemia in Bangladesh, *Journal of Data Science*, 17 (1) (2019) 195-218.
- [8] Schapire R. E., The Boosting Approach to Machine Learning: An Overview, In *Nonlinear Estimation and Classification*, 149-171, Springer, New York, 2003.
- [9] Kumar, N., Narayan Das, N., Gupta, D., Gupta, K., & Bindra, J. (2021). Efficient automated disease diagnosis using machine learning models, *Journal of Healthcare Engineering*, 1 (2021) 9983652.
- [10] Alpar R., *Applied Multivariate Statistical Methods* (Fourth Edition), Detail Publishing, 637-659, Ankara, 2013.
- [11] Beam, A. L., Manrai, A. K., & Ghassemi, M. (2020). Challenges to the reproducibility of machine learning models in health care, *Jama*, 323 (4) 305-306.
- [12] Mertler C. A., Vannatta R. A., *Advanced and Multivariate Statistical Methods: Practical Application and Interpretation*, Pyczak Publishing, Glendale, 2005.
- [13] Bland J. M., Altman D.G., The Odds Ratio, *BMJ*, 320 (7247) (2000) 1468.
- [14] Haykin S. S., *Neural Networks and Learning Machines*, Simon Haykin, 2009.
- [15] Hall M. A., *Correlation-Based Feature Subset Selection for Machine Learning*. Hamilton, New Zealand, 1998.
- [16] Witten I. H., Frank E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, *Acm Sigmod Record*, 31(1) (2002) 76-77.
- [17] [http://saedsayad.com/artificial\\_neural\\_network.htm](http://saedsayad.com/artificial_neural_network.htm), Accessed 13 June 2024