Engineering Science and Technology, an International Journal 24 (2021) 50–70

HOSTED BY

Contents lists available at ScienceDirect

Engineering Science and Technology, an International Journal

journal homepage: www.elsevier.com/locate/jestch

Classifying anemia types using artificial learning methods

Tuba Karagül Yıldız [a,*], Nilüfer Yurtay [a], Birgül Öneç [b]

[a] Computer and Information Sciences, Sakarya University, Sakarya 54187, Turkey
[b] Faculty of Medicine, Duzce University, Düzce 81620, Turkey

ARTICLE INFO

Article history:
Received 5 June 2020
Revised 23 November 2020
Accepted 1 December 2020
Available online xxxx

Keywords:
Anemia
Artificial neural network
Decision tree
Medical diagnosis
Naïve Bayes
Support vector machine

ABSTRACT

The most common blood disease worldwide is anemia, defined by the World Health Organization as a condition in which the red blood cell count or oxygen-carrying capacity is insufficient. As both a disease and a symptom, this condition affects the quality of life. Early and correct diagnosis of the type of anemia is vital in terms of patient treatment. The increasing number of patients and hospital priorities, as well as difficulties in reaching medical specialists, may impede such a diagnosis. The present work proposes a system that will enable the recognition of anemia under general clinical practice conditions. For this system, a model constructed using four different artificial learning methods. Artificial Neural Networks, Support Vector Machines, Naïve Bayes, and Ensemble Decision Tree methods are used as classification algorithms. The models are evaluated with a dataset of 1663 samples and used 25 attributes, including hemogram data and general information such as age, sex, chronic diseases, and symptoms to diagnose 12 different anemia types. Data are collected by examining patient files at a university hospital in Turkey. In addition to all the data used by the doctors, the model also utilized eight different datasets created via particular feature selection techniques. The interface is designed to provide decision support to both medical consultants and medical students. Data are classified using the four different algorithms and an acceptable success ratio is obtained for each. Each model is validated using Classification Error, Area Under Curve, Precision, Recall, and F-score metrics in addition to Accuracy values. The highest accuracy (85.6%) achieved using Bagged Decision Trees, followed by Boosted Trees (83.0%) and Artificial Neural Networks (79.6%).

1. Introduction

Anemia is the most common blood disease in the world [1]. According to the World Health Organization (WHO), anemia is a condition in which the number of red blood cells and, consequently, the oxygen-carrying capacity is inadequate to meet the body's physiological needs [2]. Anemia is also defined as a decrease in the concentration of erythrocyte mass or blood hemoglobin and hematocrit. Normal hemoglobin and hematocrit values vary according to age and sex. If hemoglobin and hematocrit values are below the threshold of normal values for the age and sex, then anemia is present. The study conducted by Kiassebaum et al. examined 189 countries, both sexes, and 20 different age groups using data and resources from the 2010 WHO study on the global burden of disease. They calculated the global anemia prevalence as 32.9%. Anemia is most commonly seen in children under five years old and in women. The most frequently encountered type of anemia is iron-deficiency anemia [3]. Since anemia, which affects the quality of life significantly, is both a disease and a symptom that accompanies many serious diseases, its treatment can be imperative in many cases, making a correct diagnosis the first step toward treatment.

The present study sought a multi-class probing solution using artificial learning architecture. The aim is to develop a system that will enable the recognition of anemia under general clinical practice conditions, as the increasing number of patients and hospital priorities, as well as difficulties in reaching medical specialists, may impede such a diagnosis. Applying this system in the primary health care services jointly with the tests required for the diagnosis of anemia will help non-specialist personnel working in these health centers. Based on this system, patients who need to be referred for treatment can be identified faster and more accurately. The 12 types of anemia most commonly encountered in a province in Turkey are classified by four different machine learning methods, with the bagged decision tree method having the highest success rate. Since there is no provision for changing the content and quality of the data, this study used a complete original dataset in

* Corresponding author.
E-mail addresses: tkaragul@sakarya.edu.tr (T. Karagül Yıldız), nyurtay@sakarya.edu.tr (N. Yurtay), birgulonec@duzce.edu.tr (B. Öneç).
Peer review under responsibility of Karabuk University.

https://doi.org/10.1016/j.jestch.2020.12.003
2215-0986/

which no numerical intervention is made. Moreover, the same methods are applied by reducing the attributes using feature selection, and the results are compared. The use of real patient data contributed significantly to the relevance of this study.

The paper is organized as follows: Section I gives an overview of the problem of diagnosing anemia. Section II presents a review of the related literature. Section III describes the anemia data used in this study in addition to defining anemia and outlining the methods used in its diagnosis. Subsequently, a summary of the well-known artificial learning methods used in the study is given and the architecture of the proposed artificial learning anemia detection system is outlined. Section IV gives the results and evaluation of the models developed in this study. Section V summarizes and discusses the results and compares them with those of previous studies. Finally, in Section VI, the motivation for this work and recommendations for possible future study topics are presented.

2. Related works

Computer-aided decision making, and analysis constitute a widespread field in the medical domain. In the present study, a system is generated to assist medical practitioners in the diagnosis of 12 different types of anemia. A review of previous studies on the classification of anemia types is carried out, along with an examination of those conducted using similar methods but with different data. Studies performed using hybrid models [4,5] are included as well.

One of the early studies of computer-aided anemia diagnosis was that of Beck et al., who designed a computer-aided system for research in medical education. They published the PlanAlyzer for diagnosis of heart disease in 1988 [6] and for anemia in 1989 [7]. This system aimed to elucidate and critique the approach of students in diagnosing a widespread medical disorder. In a study published in 1993, Lyon et al. reported that after testing and assessment of the program, it was used for seven years to teach the diagnosis of anemia and chest pain in the cardiology and hematology departments of the Dartmouth School of Medicine [8]. In 1960, Lipkin compared the data characteristics of hematological diseases with hospital data using a digital computer. For this, 49 patients and 20 diseases were selected, and the hospital data were linked to the computer program. Differential diagnoses of the hospital cases were then printed out in written form [9]. In 1976, Engle et al. introduced a computer program called HEME that provided a diagnostic analysis of 40 hematological diseases to medical consultants and was designed as a rule-based system using the Bayesian method [10].

Various algorithms developed to assist doctors in the diagnosis of iron deficiency anemia have performed successfully [11–16]. Sanap et al. devised a system for classifying the severity of anemia using complete blood count reports and the C4.5 decision tree and support vector machine algorithms with the WEKA data mining tool. They included the 10 numerical attributes of age, white blood cell count (WBC), hemoglobin (HGB), red blood cell count (RBC), hematocrit (HCT), mean cellular volume (MCV), mean cellular hemoglobin (MCH), mean cellular hemoglobin concentration (MCHC), red cell distribution width (RDW), and platelet count (PLT) and four classes of anemia types: normocytic (anemia of chronic disease), microcytic (iron deficiency and thalassemia), macrocytic (Vitamin B12 and folate deficiency), and microcytic (renal anemia). The success rate of the C4.5 decision tree algorithm was 99.42%, which surpassed the support vector machines with a success rate of 88.13% [17]. In the study conducted by Amin and Habib, the full blood count parameters of WBC, RBC, HGB, HCT, MCV, MCHC, PLT, neutrophils (NEUT), lymphocytes (LYMP), monocytes (MONO), eosinophils (EO), and basophils (BO) and the interpretation value of age were used as the data input. The classes included chronic anemia, eosinophilia, microcytic hypochromic anemia, normocytic anemia, neutrophil leukocytosis, neutrophil, unknown findings, and high erythrocyte sedimentation rate (ESR). They used the J48 decision tree, multi-layered perceptron, and Naïve Bayes as classifiers and achieved success rates of 97.16%, 86.55%, and 70.28%, respectively [18]. Iron deficiency anemia and thalassemia are two types of microcytic anemia that are at risk of being confused [19]. In a research article, a differential diagnosis of microcytic anemia was made with discriminant analysis using a training set consisting of 200 beta-thalassemia cases, 65 alpha-thalassemia cases, 170 iron deficiency anemia cases, and 45 cases having both iron deficiency anemia and beta-thalassemia [20]. Jamei and Talarposhti developed an artificial neural network (ANN) model with pattern-based input selection for iron deficiency anemia and β-thalassemia trait discrimination. This method consisted of the decision-making ability of the ANNs combined with that of a human expert. Using complete blood count results, they devised a coefficient rule base and determined the multilayer perceptron neural network input according to the calculated similarity. When compared with the performances reported by various authors using ANFIS, ANN, MLP, SVM, RBF, PNN, and KNN, their method was shown to have achieved the highest accuracy rate of 99.5% [21]. In 2015, Kishore et al. published a study using age, sex, HGB, MCV, MCH, and HCT values as input, and iron deficiency and Vitamin B12 deficiency as output. They developed a threaded ID3 approach by examining ID3 and non-threaded ID3 decision tree algorithms as methods. Using 480 data items, they tested the system with both threaded and non-threaded ID3 and Gini algorithms and reported that the method they found was usable [22].

Artificial neural networks can be used in a wide variety of areas. Yavuz et al. conducted a study for the diagnosis of iron deficiency anemia in women. Classification using ANNs and an artificial immune system (AIS) was compared with the use of KNN and the regression tree Gini algorithm. The classification performance using the Gini-based decision tree method trained by the AIS was more successful than that of the KNN method and ANNs [16]. Shaik and Subashini presented a fuzzy logic approach for anemia diagnosis. They used HGB, HCT, MCV, MCHC, WBC, reticulocyte, total iron-binding capacity (TIBC), serum iron, and hyper-segmented white cell (HSWC) laboratory test results as input parameters. As output, they used six anemia types, which included aplastic, sideroblastic, megaloblastic, chronic, myelophthisic, and iron deficiency anemias [23]. Dalvi and Vernekar conducted a study to determine the most suitable method of classifying red blood cells for anemia diagnosis. They used five ensemble learning methods (AdaBoost, bagging, stacking, voting, and Bayesian boosting) and four classifiers (k-nearest neighbor, Naïve Bayes, decision tree, and ANNs) [24]. Belginova et al. presented a rule-based approach to the diagnosis of iron deficiency anemia. They proposed a decision support system for specialist medical consultants that included patient data (e.g., identification, socio-economic status, medical history, complaints or sensations, medical indicators, and statistical information on the disease). Using these data enabled the consultant to make more accurate decisions concerning the disease [25]. Dimauro et al. conducted a study on predicting the hemoglobin value of patients using a non-invasive device capable of analyzing an image of the conjunctival region. They tested this KNN classifier on 113 individuals and obtained good results [26]. Complete blood count (CBC) testing is used to identify anemia and other hematological disorders. However, diagnosis of iron deficiency anemia and thalassemia depends on a mean cell volume (mean corpuscular volume-MCV) of<80 fl oz (fluid ounces) as an inconsistent and ambiguous feature. In a study conducted in 2005, Yeh and Cheng

proposed a solution to this problem by using the hierarchical software calculation technique of a rule-based software method. They achieved 96% accuracy on 50 samples and reported that their approach was more successful than traditional methods [27]. Allahverdi et al. published a study using the Takagi-Sugeno type neural-fuzzy (neuro-fuzzy) network method to determine childhood anemia. According to their statistical analysis, they found the errors in the system as −0.0018 MPE (mean percentage error), 0.2090 MAE (mean absolute error), 0.0511 MAPE (mean absolute percentage error), 0.2743 RMSE (root mean square error), and 0.9957 R2 (regression coefficient). They showed that the predicted anemias were very close to the measured values and reported the system to be practical and usable [28]. Maity et al. designed and developed an application to create automated anemia diagnosis reporting for acquisition and management of patient blood pathology information using the computer vision approach. The improved image processing algorithm and data mining approach could identify abnormal erythrocytes in order to analyze patient medical information. The consultant C4.5 decision tree classifier classified image samples with 98.1% accuracy and 99.6% precision [29]. Setsirichok et al., in one of their articles, proposed a classification of blood properties for a thalassemia scan via a C4.5 decision tree, a Naïve Bayes classifier, and a multi-layered sensor. The CBC properties selected were hemoglobin concentration (HBG) and mean erythrocyte volume (MCV). The average accuracy of the classification performance was found to be 93.23% and 92.60% when applying the Bayesian classifier and multilayer sensor. These results showed a combination of Naïve Bayesian classifier or multi-layer sensor with CBC and hemoglobin to be highly suitable for automated thalassemia screening [30]. In 2019, Meena et al. developed a decision support system using data mining methods for anemia in children. In their proposed model, they used the decision tree and association rules methods and obtained successful results [31]. Balaji et al. detected and diagnosed two important heart diseases, dilated cardiomyopathy (DCM) and hypertrophic cardiomyopathy (HCM), using backpropagation neural networks (BPNN) [32]. Shen et al. conducted a study in which they used a fruit-fly optimization algorithm for a parameter tuning scheme in the SVM method. They used this method on breast cancer, Pima Indian diabetes, Parkinson's disease, and thyroid datasets and stated that they had achieved successful results [33].

In 2017, Wang et al. developed a method based on the chaotic moth-flame optimization strategy for the Kernel extreme learning machine. This method performs feature selection and parameter optimization simultaneously. They successfully applied the method to Parkinson's and breast cancer datasets [34]. In 2020, Wang and Chen used the SVM method together with the whale optimization algorithm (WOA). Here, the chaotic and multiswarm algorithm improved the SVM performance of parameter optimization and feature selection. They applied their method to breast cancer, diabetes, and erythematous-squamous medical datasets and reported that they had achieved successful results [35].

An examination of studies using similar datasets revealed that they focused on diagnosing one or more general types of anemia like microcytic, normocytic, and macrocytic anemias [17,18] or thalassemia and iron deficiency anemia [19,20,21]. Table 1 summarizes the reference literature examined.

Our study diagnosed 12 different types of anemia described in the WHO International Disease Classification (ICD) Codes. In addition, the attributes used were mostly limited to a few blood parameters only. This study used 25 different attributes that an experienced medical consultant uses when diagnosing these diseases. Moreover, the data used in this study are completely original and include age, sex, chronic diseases, and symptoms as well as blood parameters.

Table 1
Review of relevant literature.

DATA	METHODS	DISEASE/CLASS	REFERENCE
MCV, MCH, MCHC, HGB, RBC	ANN, ANFIS	IDA	11. AZARKISH ET AL, 2012
HGB, MCV, SI, TIBC, FERRITIN	FFN, CFN, DDN, TDN, PNN, LVQ	IDA	12. YILMAZ, BOZKURT 2011
SERUM IRON, SERUM IRON BINDING CAPACITY, FERRRITIN	DECISION TREES	IDA	15. DOĞAN, TÜRKOĞLU, 2008
MCV, RBC, HGB, HCT, MCH, MCHT	ANN, DECISION TREES, AIS	IDA	16. YAVUZ ET AL, 2014
AGE, WBC, HGB, RBC, HCT, MCV, MCH, MCHC, RDW, PLT	DECISION TREES, SWM	NORMOCYTIC, MICTROCYTIC, MACROCYTIC, RENAL ANEMIA	17. SANAP ET AL 2011
WBC, RBC, HGV, HCT, MCV, MCHC, PLT, NEUT, LYMPH, MONO, EO, BO, AGE	DECISION TREES, MLP, NAIVE BAYES	CHRONIC ANEMIA, EOSINOPHILIA, MICROCYTIC ANEMIA, NORMOCYTIC ANEMIA, NEUTROPHIL, UNKNOWN FINDINGS, ESR	18. AMIN, HABIB, 2015
CBC	ANN	IDA, BETA THALASSEMIA	21. JAMEI ET AL, 2016
AGE, SEX, HGB, MCV, MCH, HCT	DECISION TREES (ID3, GINI)	IDA, VIT.B12 DEFICIENCY ANEMIA	22. KISHORE ET AL, 2015
RBC(IMAGES)	ADABOOST, BAGGING, STACKING, VOTING AND KNN, NAIVEBAYES, DECISION TREES, ANN	ANEMIA	24. DALVI, VERNECAR, 2016
ERITROCYTE IMAGES	DECISION TREES,	ABNORMAL ERYTROCYTE	29. MAITY ET AL, 2012
HGB, MCV	DECISION TREES, NAIVE BAYES, MULTILAYER SENSOR	THALASSEMIA	30. SETSIRICHOK ET AL, 2012
DATA OF DEMOGRAPHIC HEALTH SURVEY PROGRAM	DECISION TREES, ASSOCIATION RULES	CHILDHOOD ANEMIA	31. MEENA ET AL, 2019
ECHOCARDIOGRAM VIDEO IMAGES	BPNN, SVM	HEART DISEASES	32. BALAJI ET AL, 2016
BREAST CANCER, PIMA INDIAN DIABETES, PARKINSON'S, THYROID	SVM	BREAST CANCER, PIMA INDIAN DIABETES, PARKINSON'S, THYROID	33. SHEN ET AL, 2016
BREAST CANCER, PARKINSON'S	KERNEL EXTEME LEARNING MACHINE	BREAST CANCER, PARKINSON	34. WANG ET AL, 2017
BREAST CANCER, DIABETES, ES	SVM	BREAST CANCER, DIABETES, ES	35. WANG, CHEN, 2020
STUDENTS NATIVE PLACE	SVM, MLP	STUDENTS NATIVE PLACE IDENTIFICATION	36. VERMA ET AL, 2020
STUDENTS NATIVE PLACE	SVM, KNN, RANDOM FOREST, MLP	STUDENTS NATIVE PLACE IDENTIFICATION	37. VERMA ET AL, 2020
LEUKOCYTE IMAGES	BPNN, CNN	LEUKOCYTE CLASSIFICATION	38. BEVILACQUA ET AL, 2019
BREAST TOMOSYNTHESIS IMAGES	ANN, NON-NEURAL LEARNERS	BREAST CANCER DIAGNOSIS	39. BEVILACQUA ET AL, 2019

Notable methods in these studies using medical data included ANN, SVM, and decision tree-based methods and Naïve Bayes, KNN, and rule-based approaches. The SVM, ANN, and decision trees are successful methods that give good results and are also used in non-medical studies. For example, Verma et al. compared the SVM and MLP methods in determining the native place of students. They showed that both models give good results [36]. In their other study on the same subject, Verma et al. also used random forest and KNN methods in addition to the SVM and MLP. They stated that the random forest method gave the most successful result [37].

Hence, the literature indicates that in particular the ANN, SVM, and decision tree-based methods have been the most successful. Therefore, it was deemed appropriate to compare these methods in this study. Deep learning-based approaches are also used in medical decision-making problems. For example, in 2017, Bevilacqua et al. used feature-based backpropagation NN and deep learning-based CNN methods for the classification of leukocytes [38]. In another study conducted in 2019, they also developed a deep learning method using tomosynthesis breast images for breast cancer diagnosis. They compared optimized ANN and non-neural learner methods and used CNN for feature extraction [39]. Because our data does not contain images and the number of data items included are insufficient, deep learning methods are not used in this study. However, in the near future, we are planning to do a deep-learning-based study by increasing the size of our dataset.

3. Material and methods

In the present study, the aim is to develop a system that will enable the recognition of anemia under general clinical practice conditions. In other words, we aimed to teach the decision-making process of an experienced medical consultant to the computer program by transferring the process of diagnosing types of anemia, as the most common form of hematological diseases. For this purpose, we entered the data of patients who had presented to the hematology outpatient clinic with a pre-diagnosis of anemia into the computer program and evaluated the examinations carried out on the patients with anemia-related complaints. State-of-the-art artificial learning methods used in the learning phase. After the learning process, we tested the system using new data, and investigated its ability to make decisions in the same way as an experienced medical consultant. At this stage, we used the ROC analysis method. The main outcome of the study is the transference of the decision-making method used by an experienced medical consultant. Another outcome is the providing of decision support to doctors and medical students. Moreover, this system can also carry out patient follow-up procedures.

To determine the presence of anemia, first, the HGB value is examined by the expert, as seen in Fig. 1. In the next step, the MCV value is examined. If the MCV is<80, then the anemia type is microcytic. If the MCV is between 80 and 100, then the anemia type is normocytic. If the MCV is more than 100, then the anemia type is macrocytic. After the first phase of identification, the expert may require further investigations and advanced tests for a definite diagnosis. The detailed types and/or causes of anemia are shown in Fig. 1.

In order for the computer to diagnose anemia like an expert medical consultant, real patient data and the advice of an experienced medical specialist are needed. This specialist provided information on the features required and methods to be followed in the diagnosis of anemia. Furthermore, the data required could only be obtained through the approval of the Ethics Committee. Once ethical approval is obtained, the data are transferred from the hospital database to the program interface, as shown in Fig. 2. This interface is based on the opinion of the experienced medical specialist.

Thus, as seen in Fig. 2, data from the interface are processed by classifiers and the results are interpreted. Since the aim is to make decisions in the same way as the experienced medical specialist would decide, care is taken not to make any qualitative changes in the data. The structure of the proposed method can be seen in Fig. 3.

After the data are obtained using the program interface, four basic models are developed for the classification process: support

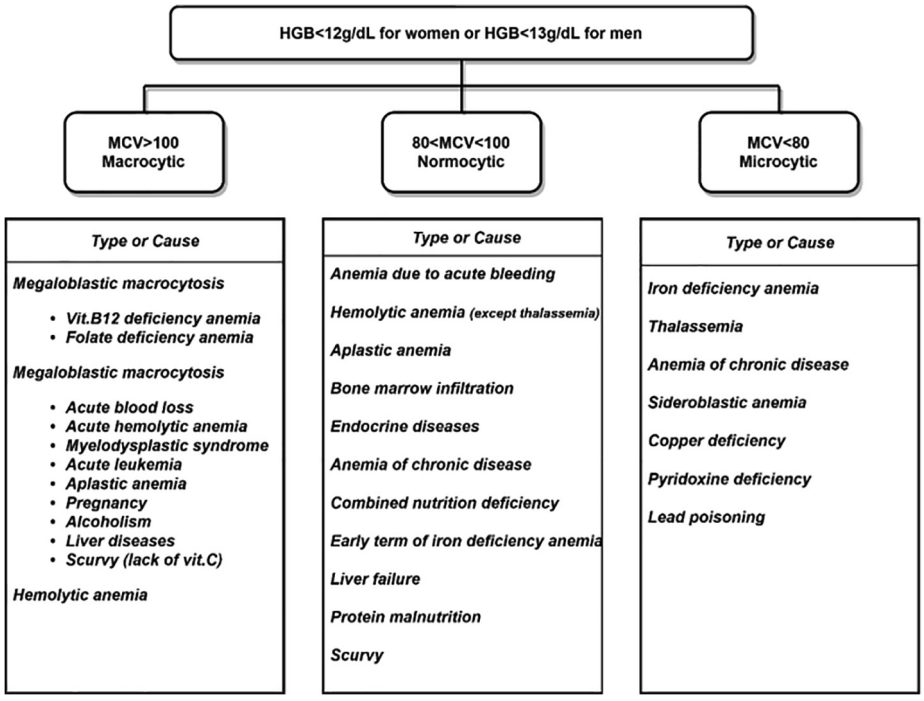

Fig. 1. Classification of anemia according to erythrocyte morphology [40].

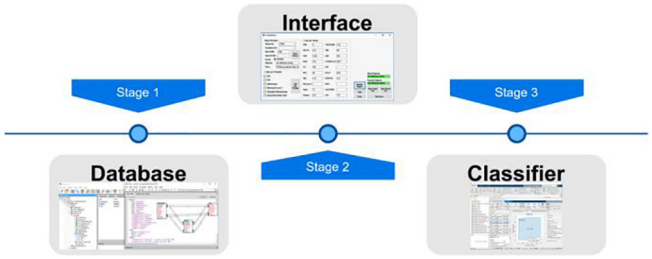

Fig. 2. Program interface used to provide data from clinical database.

vector machines (SVM), decision trees (DT), artificial neural networks (ANNs), and Naïve Bayes. These models are selected because they are state-of-the-art classification methods and are expected to give promising results. Finally, the developed model is recorded and tested on new data and a performance evaluation is carried out using receiver operating characteristic (ROC) analysis [41]. Besides each model is validated using classification error, AUC, precision, recall, and F-score metrics in addition to accuracy values.

3.1. Dataset

The data used in the study are actual patient data obtained from the Düzce University Research and Application Hospital with permission from the Hospital Ethics Committee. In accordance with the Turkish law for the protection of personal data, the ethics committee had to be informed as to the kind of data we wanted to use. Therefore, we consulted an experienced medical specialist and determined the attributes she uses to diagnose anemia types. All the attributes used in the dataset are shown in Table 2.

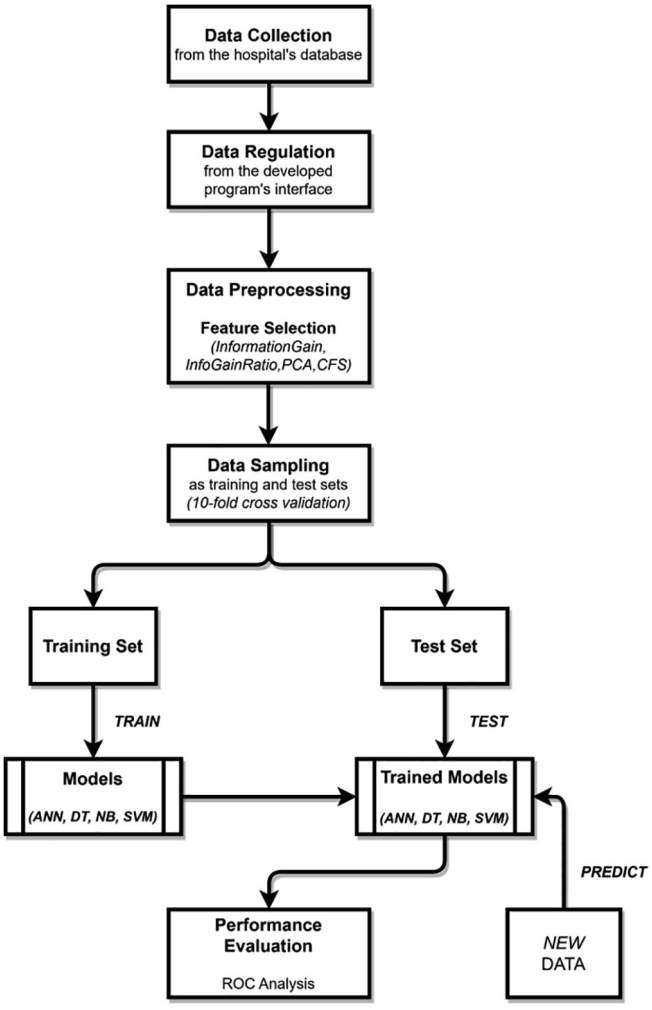

Fig. 3. Flow chart of the present work.

In addition to those seen in Table 2, there are also other attributes in the raw dataset that enabled us to organize our data. The archive number is a unique attribute used to identify a patient in the hospital records. From the admission number of the patient and the approval date, we could determine how often the same patient had applied to the clinic. Other Information consisted of patient histories not included as attributes for this study. Our data consisted of only 30 attributes. As explained above, four of them are not used. Information such as patient age, sex, and the presence of symptoms and chronic diseases are attributes that play an important role in determining anemia type. Bilirubin values in the blood analysis are used to assess liver and gall bladder function. The C-reactive protein (CRP) provides information about the presence of inflammation in the body. Iron values in the blood are used in the evaluation of all types of anemia, iron deficiency, and iron poisoning. The ferritin value is used in the diagnosis of iron deficiency anemia, chronic disease anemia, and thalassemia and is also important for monitoring iron-loading treatment. Folate refers to the folic acid value in the blood and is used in the evaluation of megaloblastic and macroscopic anemia as well as being used to monitor the treatment of folate deficiency anemia. The hematocrit (HCT) shows the amount of hemoglobin and erythrocytes present in the blood. The hemoglobin (HGB) shows the total amount of hemoglobin present in blood and is the first value that indicates anemia in an investigation of complete blood count parameters. The creatinine value in the blood is used in the evaluation of kidney function. The mean cellular hemoglobin (MCH) shows the total amount of hemoglobin in the erythrocytes. The mean cellular hemoglobin concentration (MCHC) is the percentage of hemoglobin concentration in the erythrocytes. The mean corpuscular volume (MCV) is the average size of the red blood cells carrying oxygen. The NEUT is the number of neutrophils in the blood and the PLT is the number of platelets, whose function is to enable blood to clot. The red blood cell count (RBC) is the number of erythrocytes present in the blood and the red cell distribution width (RDW) shows the distribution width of the erythrocytes in the blood. The total iron-binding capacity (TIBC) and unbound iron-binding capacity (UIBC) are also important parameters used to diagnose anemia types. Vitamin B12 is an essential vitamin for hematopoiesis and normal neuronal functions. In the case of low vitamin B12, vitamin B12 deficiency anemia may be considered. The white blood cell count (WBC) is the number of leucocytes in the blood. These act as the body's defense and are responsible for the immune system [40,42,43].

There are 1663 data ıtems in the dataset. The distribution of these data according to diagnosis is given in Table 3. The distribution of diseases associated with anemia is irregular and unbalanced. There were 1109 female and 554 male patients in our dataset. Women are known to have a high prevalence of anemia and these data confirm this situation. Iron deficiency anemia, constituting 21% of the dataset, is the most common type of anemia seen in the region, while the least common is the thalassemia trait.

However, it should be kept in mind that to obtain these data for use in our study, the ICD codes are limited to between D50 and D64.9 and the attributes are limited to 30 different features. Therefore, the attributes needed to diagnose anemia-associated diseases are selected on the recommendation of the experienced medical specialist.

This study is conducted in Düzce province, Western Black Sea Region of Turkey. The anemia types listed here are the 12 most

Table 2
List of attributes in the dataset.

Attribute Name	Type	Min	Max	Avg
Age	Numeric	20	109	55.4
Sex	Numeric	0	1	0.6
Chronic Disease	Numeric	0	1	0.6
Symptoms	Numeric	0	1	0.5
CRP (C Reactive Protein)	Numeric	0	27.6	1.2
D. Bilirubin (Direct Bilirubin)	Numeric	0	11.3	0.2
Iron	Numeric	4	377	77.8
Ferritin	Numeric	0	2338.4	166
Folate	Numeric	1	99.6	11.7
HCT (Hematocrit)	Numeric	11	64.5	35.4
HGB (Hemoglobin)	Numeric	1	22.9	11.6
I. Bilirubin (Indirect Bilirubin)	Numeric	−0.1	5.01	0.4
Creatinine	Numeric	0.2	8	0.9
MCH (Mean Cellular Hemoglobin)	Numeric	13.9	45.2	27.4
MCHC (Mean Cellular Hemoglobin Concentration)	Numeric	25.6	38.2	33.1
MCV (Mean Cellular Volume)	Numeric	49	126.6	82.7
NEUT (Neutrophil Count)	Numeric	0	47.8	3.9
PLT (Platelet Count)	Numeric	1	1239	260.7
RBC (Red Blood Cell Count)	Numeric	1.2	45.2	4.3
RDW (Red Cell Distribution Width)	Numeric	11.2	38.2	17.2
T. Bilirubin (Total Bilirubin)	Numeric	−0.02	5.7	0.7
TIBC (Total Iron-Binding Capacity)	Numeric	104	697	353.8
UIBC (Unbound Iron-Binding Capacity)	Numeric	−9	676	273.1
Vitamin B-12	Numeric	13.1	2000	512.4
WBC (White Blood Cell Count)	Numeric	0.7	431.3	7.6
Diagnosis	Polynomial	–	–	–

Table 3
Codes and distribution of diagnoses in the dataset.

ICD-10 Codes	Diagnosis	Count	%
D64	Anemic	123	7.39
–	Non-Anemic	184	11.06
D50	Iron Deficiency Anemia	351	21.10
D50-D52	Iron and Folate Deficiency Anemia	187	11.24
D50-D51	Iron and Vit. B12 Deficiency Anemia	164	9.86
D52	Folate Deficiency Anemia	234	14.07
D51-D52	Folate and Vit. B12 Deficiency Anemia	55	3.30
D59	Hemolytic Anemia	42	2.52
D63	Anemia of Chronic Disease	170	10.22
D56	Thalassemia	80	4.81
D57	Thalassemia Trait	23	1.38
D51	Vitamin B12 Deficiency Anemia	50	3.006

common types of anemia in the province. As the data are taken from a hospital, all the patients in the dataset had at least one hematological disorder. The "non-anemic" patient group did not consist of healthy individuals and for this reason, is not considered as a control group. Since they are suffering from hematological disorders outside the anemia group, the use of the term "non-anemic" is considered appropriate.

This study used the data of patients who had applied to the Hematology Outpatient Clinic at Düzce University Research and Application Hospital for whom anemia and related diseases (ICD codes D50.0 –D64.9) are entered as the diagnosis or pre-diagnosis. The patient data used included: age, sex, chronic disease, symptoms, CRP (C reactive protein), D. bilirubin (direct bilirubin), iron, ferritin, folate, HCT (hematocrit), HGB (hemoglobin), I. bilirubin (indirect bilirubin), creatinine, MCH (mean cellular hemoglobin), MCHC (mean cellular hemoglobin concentration), MCV (mean cellular volume), NEUT (neutrophil count), PLT (platelet count), RBC (red blood cell count), RDW (red cell distribution width), T. bilirubin (total bilirubin), TIBC (total iron-binding capacity), UIBC (unbound iron-binding capacity), vitamin B-12, and WBC (white blood cell count).

Among the data of anemia-related diseases, the hemogram had to be evaluated first. The interpretation of the hemogram is based on the WHO definitions of anemia [2] and the recommendations of the Hematology Laboratory Guide [44] published by the Turkish Society of Hematology (TSH) in October 2014. According to these recommendations, when examining the data of a patient, the hemoglobin values are considered first, with HGB < 13 g/dL in male patients and HGB < 12 g/dL in female patients described as anemia. Following that, patients had to be classified as microcytic, normocytic, or macrocytic according to the MCV value. The ferritin value of those with microcytic anemia (MCV < 80) is then questioned, and iron deficiency or thalassemia diagnoses noted accordingly. In patients who are not suspected of iron deficiency (ferritin greater than 15 according to TSH anemia guidelines), iron is assessed according to iron-binding capacity. With this evaluation, we aimed for the differential diagnosis of iron deficiency or chronic disease anemia. In each patient with anemia, vitamin B12 and folic acid values also had to be evaluated and determined as vitamin B12 or folate deficiency anemias accompanying other anemia types (iron deficiency, chronic disease anemia, thalassemia, etc.) or especially as macrocytic-defined anemias. In addition, other series (white blood cells and platelets) of patients with anemia had to be evaluated and if these are not normal (either high or low values), a peripheral smear had to order. According to WHO criteria, any symptoms and findings that might require urgent transfusion in those with severe anemia should receive immediate attention. In addition, if there is significant evidence for the etiology of anemia, it is vital to be on the alert for each type.

Anemia Types and Diagnostic Criteria:

Anemias constitute the most common blood disease group in the world as well as in Turkey [1]. According to WHO, anemia is a condition in which the number of red blood cells (and, accordingly, their oxygen-carrying capacity) is insufficient to meet the physiological needs of the body [2]. Anemia is also defined as a decrease in erythrocyte mass or blood hemoglobin and hematocrit concentration. Normal hemoglobin and hematocrit values vary according to age and sex. Anemia is present when hemoglobin and hematocrit values are below the lower limit of normal values for the age and sex.

The main causes of anemia include a deterioration in the morphological (structural) and/or physiological functions of the erythrocytes. Anemias can occur for four main reasons:

1. Erythrocyte production disorder (insufficient erythrocyte production by bone marrow)
   a. Bone marrow malfunction, bone marrow failure (e.g., aplastic anemia and infection-, drug-, or cancer-related bone marrow failure)
   b. Impairment of erythropoietin synthesis, 90% of which is released from the kidneys and plays a very important role in the ripening of erythrocytes (e.g., chronic kidney failure, hypothyroidism, and rheumatic diseases)
2. Structural and functional impairment of erythrocyte maturation (e.g., iron deficiency, hemoglobin structure and function disorders, lead poisoning, vitamin B12 deficiency, and folic acid deficiency)
   a. Early destruction of erythrocytes (hemolytic anemias)
   b. Causes of erythrocyte destruction (e.g., erythrocyte membrane disorders, erythrocyte enzyme deficiency, and hemoglobinopathies)
3. Non-erythrocyte causes (e.g., immune, and non-immune causes)
4. Blood loss (hemorrhaging)

Common clinical indications of anemia are weakness, fatigue, and paleness. Bone and joint pain, enlarged lymph nodes, and enlarged liver and spleen can be observed in leukemia and some other hematological diseases. There may be symptoms like palpitations, headaches, frequent infections, impaired nails, loss of appetite, loss of taste, painful tongue, sores in the mouth, and the desire to eat non-food substances like soil, cement, or ice (pica). Patients with long-term anemia can tolerate anemia symptoms more comfortably and may not have significant complaints [2,40,45,46].

The first laboratory tests to be requested in the patient with anemia are complete blood count and erythrocyte indices including MCV (mean erythrocyte volume), MCH (mean erythrocyte hemoglobin), MCHC (mean erythrocyte hemoglobin concentration), and RDW (erythrocyte distribution width).

Initial speculations as to the cause of anemia are obtained through the patient history, physical examination, and test results. Later, additional tests can be ordered for a definitive diagnosis [45]. The classification of anemia according to erythrocyte morphology is as shown in Fig. 1.

Production disorder and, hypo-proliferative anemias are characterized by a low reticulocyte production index and little or no change in erythrocyte structure. Damage to the premature stem cell pool of the bone marrow structure can occur as a result of erythropoietin impulse or iron deficiency. Erythropoietin is a glycoprotein hormone that acts as a cytokine (a group of proteins that enable cells to communicate with each other) for erythrocytes. It is produced in the kidneys and is the hormone responsible for the control of erythrocyte production [46].

In ripening disorders, a low reticulocyte production index is accompanied by a macrocytic or microcytic erythrocyte structure. Impairment of erythrocyte precursor cell ripening order may be due to folic acid and vitamin B12 deficiency, chemotherapy, or a myelodysplastic or preleukemic condition. Because these are all associated with nuclear maturation disorders, patients may have macrocytic anemias, megaloblastic bone marrow structure, and varying degrees of infectious erythropoiesis.

Patients with increased hemolysis-related erythrocyte destruction exhibit an increase of more than triple the normal balanced reticulocyte index level and an erythrocyte structure that may or may not be differentiated.

The first step when classifying anemia is important for both diagnosis and treatment. Treatment of the disease will also vary according to functional impairment [43,47].

In the WHO disease classification guide, iron deficiency anemia is included in the dietary anemia group (ICD codes D50-53) together with vitamin B12 deficiency anemia and folate deficiency anemia. Thalassemia, thalassemia trait, and hereditary and acquired hemolytic anemias due to enzyme disorder are in the hemolytic anemia group (ICD codes D55-59). The group of aplastic and other anemias (ICD codes D60-64) includes aplastic anemia, chronic disease anemia, and other anemias [48]. However, when classified according to erythrocyte morphology, iron deficiency anemia is included in the microcytic anemias, and vitamin B12 and folate deficiency anemias are included in the macrocytic anemias. Moreover, situations where these diseases are seen together were prevalent in the clinic. According to the classification given in Fig. 1, iron deficiency anemia, thalassemia, and thalassemia trait are included in the microcytic anemias. Although some chronic disease anemias are microcytic, most of them are included among the normocytic anemias. Hemolytic anemias are also included in normocytic anemias. Vitamin B12 deficiency and folate deficiency anemias are included in macrocytic anemias.

Iron deficiency anemia is the most common type of anemia. It is the protein-containing iron structure in the center of hemoglobin that allows red blood cells to transport oxygen from the lungs to the tissues. This function cannot take place when body iron is lost, and consequently, various symptoms such as weakness, fatigue, and shortness of breath are seen. It is most common in women and children. Measurement of MCV, iron, ferritin, and iron-binding capacity values is important in the diagnosis. In addition, the possibility of internal bleeding should also be eliminated. Iron deficiency anemia can be treated with iron supplements and a diet containing iron-rich foods [49,50].

Vitamin B12 plays an important role in red blood cell production and the functioning of the nervous system. When the vitamin B12 in the body is insufficient, healthy production and division of red blood cells cannot be carried out. As a result, problems occur with the passing of the RBCs from the bone marrow to the blood, causing various bodily symptoms. The HGB, MCV, and vitamin B12 values are important measurements in the diagnosis. Vitamin B12 deficiency can be treated with adequate nutrition and vitamin B12 supplementation [49,51].

Folic acid is a substance found in fruits, green leafy vegetables, and meat. Deficiency occurs when its intake is inadequate, or it is not sufficiently absorbed by the body. The serum folate level is an important measurement in the diagnosis. Folic acid deficiency can be treated with a folic acid-rich diet and supplements [51,52].

Chronic disease anemia is a type of anemia that accompanies chronic diseases such as cancer and diabetes, heart, kidney, and rheumatic diseases, infections, and inflammation, especially in older individuals. Low levels of serum iron and total iron-binding capacity are important measurements in the diagnosis. For the treatment of this type of anemia, the underlying disease must be treated first [53].

Thalassemia is a disease that occurs when few or no hemoglobin chains can be produced. It is a genetic transition disease, and therefore, the heterozygotes become carriers and the homozygotes become ill. The HGB, HCT, erythrocyte count, and MCV, MCH, and MCHC index values are important measurements in the diagnosis. The transfusion is administered in the treatment of patients, although not usually in the case of carriers. Thalassemia patients must be observed throughout their entire life [53].

Hemolytic anemia can be defined as a condition in which the red blood cells are destroyed at a faster rate than they are produced. The reason may be hereditary or acquired. Although the patient's history is important in diagnosis, laboratory methods

such as complete blood count and peripheral smear, hemoglobin electrophoresis, and bone marrow tests are also used. Medication, surgical intervention, blood transfusion, and marrow and stem cell transplantation can be applied in its treatment [49,53].

In our study, patients diagnosed with anemia are also diagnosed with anemias other than those described above. In the WHO definition, the ICD 10 code D64 is included as "other anemias". The complete blood count and especially the HGB, HCT, and RBC values are used in the diagnosis of anemia. Since it is a disease that significantly affects the quality of life, it is important to recognize and treat the anemia type [43]. Ultimately, the classification is performed with the 25 attributes included in Table 1. Since the data are taken from the patient files individually, no null value is included. Only the digitization and normalization of the data are performed in the pre-processing.

3.2. Classifiers

In the present study, the performance of well-known classification methods evaluated by creating a completely original dataset. As methods widely used in the literature, ANNs, support vector machines, decision trees, and Naïve Bayes are chosen as the classifiers. These are state-of-the-art classification methods that give promising results. In addition, these methods have been found to produce successful results when used with medical data. Therefore, these methods are applied and the results are compared. The classification process is performed using the MATLAB® R2020a version.

3.2.1. Artificial neural networks (ANNs)

The purpose of this study was to enable the computer to perform the diagnostic procedure in the same way as it is performed by doctors. The ANN is a state-of-the-art method that roughly models the learning process of the human brain and was considered to be a suitable method within the scope of this study. Just as the human brain learns by analyzing samples, artificial neural networks learn from samples as well. An ANN consists of interconnected cells ("neurons") like the nerve cells in the human brain. In the ANN model, each neuron must have inputs, weights, addition and activation functions, and outputs [54]. Each input has a weight that affects the activation level of the neuron. The output value is reflected in the transfer function as the sum of the input signals multiplied by the weights. The learning capacity of an artificial neuron is determined by regulating the weights of the selected learning algorithm [55]. In the ANN method, initially, a training set is created and both inputs and outputs are given to the network. The outputs produced by the network are then compared with the actual outputs. After the error calculation is made, the weights are updated, and this process iterated until the lowest error rate is reached and the training process accomplished. In the next step, the model created in the training process is run again with a test set preferably consisting of different data, and the learning of the network is tested. A two-layered feed-forward neural network model was used for this study. The basic structure of the proposed neural network is presented in Fig. 4.

There were 1663 samples and 25 features in the database, as introduced in the previous section. Therefore, the input layer of the ANN also consisted of 25 neurons. As an output, each sample belonged to one of the 12 different classes. Furthermore, the neural network model had 10 hidden layers, with each layer made up of 50 neurons. The sigmoid transfer function was selected as the activation function. Equation (1) shows the neural network's sigmoid transfer function, where $\times$ indicates inputs and $f(x)$ indicates output.

$$f(x) = \frac{1}{1+e^{x}} \tag{1}$$

The neural network model designed for this study was a two-layer feed-forward neural network, with sigmoid functions in the hidden layer and softmax functions in the output layer. The training process was performed using a scaled conjugate gradient backpropagation algorithm.

Nearly 60% of the dataset (997 samples) was used for the training process, 20% (333 samples) for the testing process, and 20% (333 samples) for the validation process.

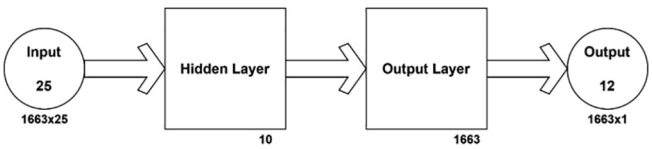

Fig. 4. Structure of the proposed neural network algorithm.

3.2.2. Support vector machines (SVM)

Support vector machines are among the supervised learning methods that can be applied to classification or regression problems. The classification is performed by dividing the input space of the dataset linearly or non-linearly. The linear decision line is drawn so that the traced samples have a minimum distance between each other, but maximum line spacing. It is a method that gives good results in real-world applications [55]. The structure of the SVM method is seen in Fig. 5.

The calculation for the hyperplane (H) is given in Equation (2), where w indicates a set of weights, b indicates bias, and x indicates input sample features.

$$H : w.x_i + b = 0 \tag{2}$$

In the SVM method, the kernel function is one of the important parameters for classifier success. For this study, three different kernel functions were used: linear, cubic, and quadratic. The kernel scale was selected automatically by MATLAB. Data regularization and standardized parameters were also set to represent true properties.

3.2.3. Decision tree (DT)

The decision tree classification process is a method of testing whether a feature can be distinguished in the classes in a dataset. Each feature found forms a branching condition of the tree. With this method, all data in the dataset are intended to be placed in

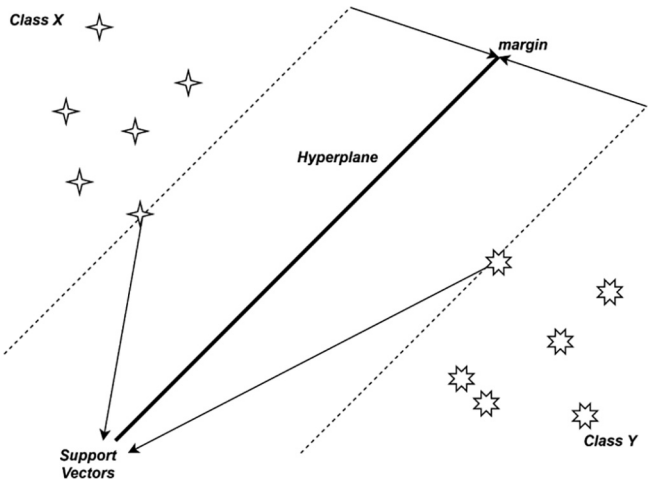

Fig. 5. Structure of the SVM.

one of the classes and in this way, a class definition is made at the same time. The results are easy to understand and interpret [55]. This method also gives good results and is commonly used with medical data. The structure of the decision tree is seen in Fig. 6.

The decision tree method used for this study was carried out using two different ensemble methods: boosting and bagging. Ensemble techniques are used in the solution of multi-class problems. The goal is to improve performance by grouping the binary classes to form multi-classes. In the AdaBoost method, at each iteration, the weights of misclassified samples of the decision tree are increased and the weights of correctly classified samples are reduced. In the subsequent iterations, the updated weights are used. Thus, the algorithm concentrates on misclassified items. In the bagging method, the data is divided into subsets and a learning model is applied for each. Bagged trees are created by combining a plurality of decision trees and can give successful results where the classes are categorical and nonlinear. In this study, a different decision tree was created for each subset of the dataset. Accuracy was calculated by taking account of the average performance of each tree [56].

3.2.4. Naïve Bayes

Naïve Bayes is a machine learning algorithm based on Bayes' theorem. In Naïve Bayes, when the class each data item belongs to is clear, the goal is to create a rule that will determine the class label of the next data item to arrive [57]. This circumstance is also called conditional probability. It is based on the principle of the value taken by the relevant attribute when the data class is labeled. When applied to a dataset, this is expressed as in Equations (3) and (4).

$$P(PC|S) = \frac{PS|PC) * P(PC)}{P(S)} \tag{3}$$

$$\mathrm{P(PC|S)} = \mathrm{P(S|PC)} * \mathrm{P(PC)} \tag{4}$$

Where,
PC: Parent Category,
S: Successful,
P (PC): The probability of the parent category,
P (S): The probability of the class label being successful,
P (PC | S): The probability of parent category when the class label is successful,
P (S | PC): The probability that the class label is successful in the case of parent category.

3.3. Feature selection

The objective of this study was to transfer the decision-making process of the doctor to the computer. Therefore, it was necessary to use the same data used by the doctor. However, the computer learning process is done using various artificial learning methods.

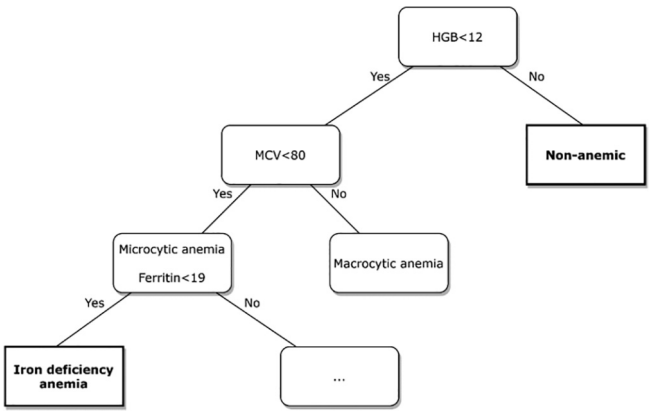

Fig. 6. Decision tree structure.

In artificial learning methods, the effects of the attributes given to algorithms on the result are an important factor, and thus, the use of feature selection methods is essential. When selecting attributes, the basic process involves determining the weights for each attribute and eliminating the attributes according to their weights. The weight of an attribute is usually calculated in the range (0,1) or (-1, +1). The closer the weight value is to 1 or −1, the more important its effect will be on the result. If this value is close to 0, it does not have much effect on the result and the attribute can be eliminated. Numerous methods were used in this study for the selection of attributes. These include information gain, information gain ratio, principal component analysis (PCA), and correlation-based attribute subset (CFS) selection. The WEKA and RapidMiner data mining tools were used for this task. Information gain, information gain ratio, and PCA-based feature selection processes were performed using RapidMiner. Various methods were tried using WEKA, but other than CFS, the methods selected all 25 attributes and thus, the attributes could not be reduced. By applying these four methods, eight different datasets were obtained in addition to the original dataset. The attributes included in each dataset can be seen in Table 4. Consequently, the classification algorithms were run for a total of nine datasets and the results were compared.

3.3.1. Information gain

The information gain of a feature in a dataset is the ability to determine the class to which it belongs. For example, if the value of an attribute in a dataset enables us to know the class to which it belongs, the information gain of that attribute will be 1. Otherwise, if the value of an attribute gives us no information about its class, then the information gain of that attribute will be 0. Essentially, in order to understand information gain, the theory of entropy must be understood. Entropy can be defined simply as the information contained in the data. Shannon's entropy formula is given in Equations (5) and (6) [58].

$$E(class) = -\sum_{i=1}^{c} p_i log_2(p_i) \tag{5}$$

$$E(class, attribute) = -\sum_{j=1}^{v} \frac{C_j}{C} E(Cj) \tag{6}$$

Here, $\boldsymbol{c}$ is the number of classes (the number of values the target variable can take), $\boldsymbol{p_i}$ is the probability that a random data is from class $\boldsymbol{i}$, $\boldsymbol{j}$, $\boldsymbol{v}$ is the number of attributes (the number of values the predictive variable can take), and $\boldsymbol{C}$ represents the class values.

If an attribute in the dataset has a different value for each class, its entropy is 0. In other words, the class can be determined according to the value of that property and there is no need to look at other properties. In this case, the information gain is 1. The less correlated a feature is with its class value, the lower the information gain will be. The equation used for the information gain calculation is given in Equation (7).

$$InformationGain(class, attribute) = E(Class) - E(Class, Attribute) \tag{7}$$

The Class and Attribute are represented accordingly in this equation. After dividing the dataset into classes, the information gain is obtained by subtracting the entropy value of the determined attribute from the entropy value of all classes. In entropy, the importance of the attribute decreases as its value approaches 1, whereas, in information gain, the importance increases as its value approaches 1.

The information gain values obtained for the present study are shown in Table 5 and Fig. 7. Three different datasets were created by selecting three attributes with a weight value greater than 0.5,

Table 4
Datasets created by feature selection.

Attributes	CFS	PCA	Information Gain			Information Gain Ratio		
			>0,1	>0,2	>0,5	>0,3	>0,4	>0,5
Age	✓	✓	✓					
Sex								
Chronic Disease								
Symptom								
CRP		✓						
D.BILIRUBIN						✓	✓	
IRON		✓	✓			✓	✓	
FERRITIN	✓	✓	✓	✓	✓	✓	✓	✓
FOLATE	✓	✓	✓	✓	✓	✓	✓	✓
HCT		✓	✓			✓	✓	
HGB	✓	✓	✓	✓		✓		
I.BILIRUBIN			✓			✓	✓	✓
CREATININ								
MCH		✓	✓			✓	✓	
MCHC						✓		
MCV	✓	✓	✓					
NEUT#						✓		
PLT		✓						
RBC		✓	✓			✓	✓	✓
RDW		✓				✓		
T. BILIRUBIN	✓					✓	✓	✓
TIBC		✓	✓			✓	✓	✓
UIBC		✓	✓	✓		✓		
VITAMIN B12	✓	✓	✓	✓	✓	✓	✓	✓
WBC		✓						

five attributes with a weight value greater than 0.2, and 13 attributes with a weight value greater than 0.1.

3.3.2. Information gain ratio

The information gain ratio is the ratio of information gain to intrinsic value. Basically, intrinsic value is the amount of information needed to identify the class to which a data item belongs. In cases where the information gain is not sufficient, the value of the gain ratio can be applied. Equation (8) shows the calculation of the information gain ratio.

$$InformationGainRatio(class, attribute) = \frac{E(class) - E(class, attribute)}{E(attribute)} \quad (8)$$

Table 5
Attribute weights by information gain.

Attribute	Weight
Symptom	0,000
NEUT#	0,014
MCHC	0,048
CREATININ	0,048
CRP	0,055
RDW	0,062
WBC	0,072
Sex	0,087
PLT	0,088
Chronic Disease	0,089
D.BILIRUBIN	0,102
T. BILIRUBIN	0,103
Age	0,105
I.BILIRUBIN	0,105
MCV	0,123
RBC	0,124
IRON	0,131
MCH	0,133
HCT	0,184
TIBC	0,197
UIBC	0,244
HGB	0,257
VITAMIN B-12	0,694
FOLAT	0,813
FERRITIN	1,000

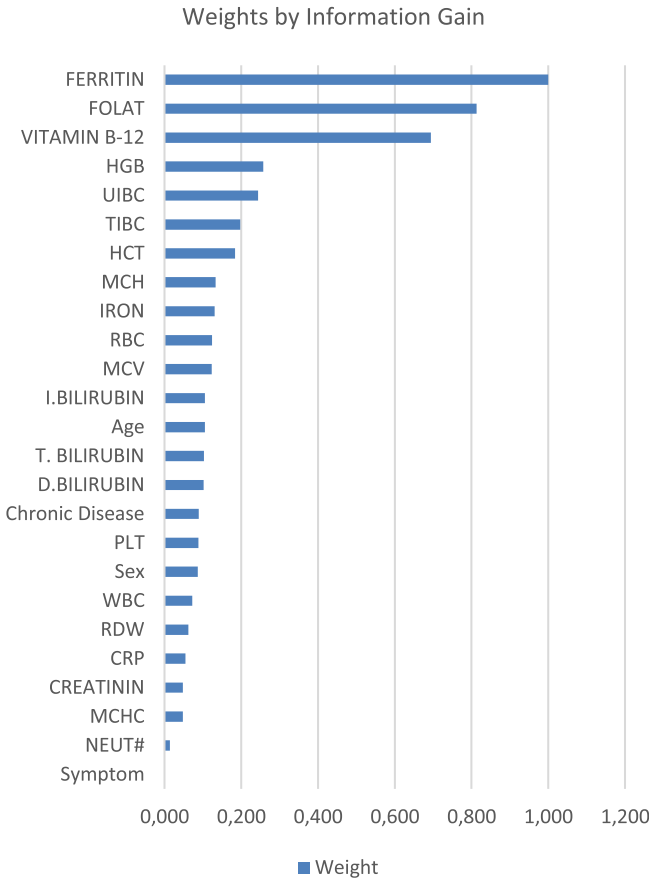

Fig. 7. Attribute weights by information gain.

In the information gain ratio calculation, as the value of the attribute approaches 1, its importance increases [59]. The information gain ratio values obtained for the present study can be seen in Table 6 and Fig. 8. Seven attributes with a weight value greater

than 0.5, 11 attributes with a weight value greater than 0.4, and 16 attributes with a weight value greater than 0.3 were selected, thus creating three different datasets.

3.3.3. Principal component analysis (PCA)

The principal component analysis is a frequently used feature selection method. To perform PCA, first of all, the relationship between attributes must be determined. In the utilization of classification algorithms, having a large number of attributes that are related to each other is undesirable. Having independent attributes in the dataset ensures stronger results in the classification process. Using PCA, numerous attributes correlated with each other are represented by fewer attributes with no correlation.

In the weight calculation made with PCA, the importance of the attribute increases as the value approaches 1. The weight values obtained for the present study are shown in Table 7 and Fig. 9. With PCA, 16 attributes with a weight value other than 0 were selected and a new dataset was created.

Table 6
Attribute weights by information gain ratio.

Attribute	Weight
Symptom	0,000
Sex	0,095
Chronic Disease	0,097
MCV	0,263
PLT	0,304
CREATININ	0,309
WBC	0,310
Age	0,314
CRP	0,320
RDW	0,373
HGB	0,385
MCHC	0,417
NEUT#	0,417
UIBC	0,457
HCT	0,485
MCH	0,485
IRON	0,495
D.BILIRUBIN	0,506
I.BILIRUBIN	0,594
TIBC	0,598
RBC	0,609
T. BILIRUBIN	0,621
FOLAT	0,872
VITAMIN B-12	0,990
FERRITIN	1,000
Attribute	Weight
Symptom	0,000
Gender	0,095
Chronic Disease	0,097
MCV	0,263
PLT	0,304
CREATININ	0,309
WBC	0,310
Age	0,314
CRP	0,320
RDW	0,373
HGB	0,385
MCHC	0,417
NEUT#	0,417
UIBC	0,457
HCT	0,485
MCH	0,485
IRON	0,495
D.BILIRUBIN	0,506
I.BILIRUBIN	0,594
TIBC	0,598
RBC	0,609
T. BILIRUBIN	0,621
FOLAT	0,872
VITAMIN B-12	0,990
FERRITIN	1,000

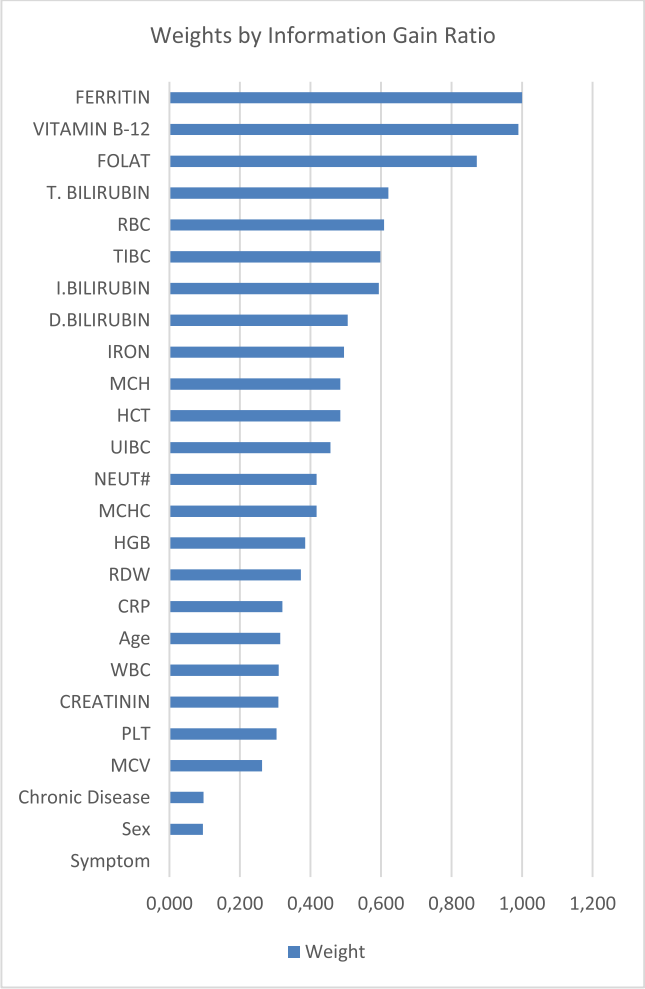

Fig. 8. Attribute weights by information gain ratio.

3.3.4. Correlation-based feature subset selection (CFS)

This method aims to choose a set of attributes that can be useful for classification. For an attribute to be effective, there should be a high correlation of that attribute with the class, while it should be less correlated with other attributes. Each attribute set is considered separately, and its correlation weight value is calculated. The subset with the highest weight from the retrieved subsets is presented to the classification algorithm. [60] The attributes in the subset obtained for the present study can be seen in Fig. 10. A total of 178 subsets were evaluated and the weight value of the best subset was found to be 0.579. In this subset, seven attributes (Age, Ferritin, Folate, HGB, MCV, T. Bilirubin, and Vitamin B12) were selected and a new dataset was created.

3.4. Evaluation

For all methods, 10-fold cross-validation is used in this study. In the cross-validation method, the dataset is divided into 10 different subsets. When a group is a test set, the remaining nine groups are used as training sets in turn. In this way, all the combinations are tested so that each of the 10 datasets is a test set once, and a performance value is found by taking the average of each result.

Receiver operating characteristic (ROC) analysis is used for performance measurement in this study. The ROC analysis is an effective method for measuring the performance of machine learning and data mining techniques [41,61]. The confusion matrix for

Table 7
Attribute weights by PCA.

Attribute	Weight
UIBC	−0.085
TIBC	−0.061
PLT	−0.009
HCT	−0.002
HGB	−0.001
RBC	−0.001
Sex	−0.000
Symptom	0.000
I.BILIRUBIN	0.000
MCHC	0.000
CREATININ	0.000
D.BILIRUBIN	0.000
T.BILIRUBIN	0.000
Chronic Disease	0.000
NEUT#	0.000
CRP	0.001
FOLATE	0.001
RDW	0.001
MCH	0.001
WBC	0.002
MCV	0.004
Age	0.013
IRON	0.024
FERRITIN	0.322
VITAMIN B12	0.941

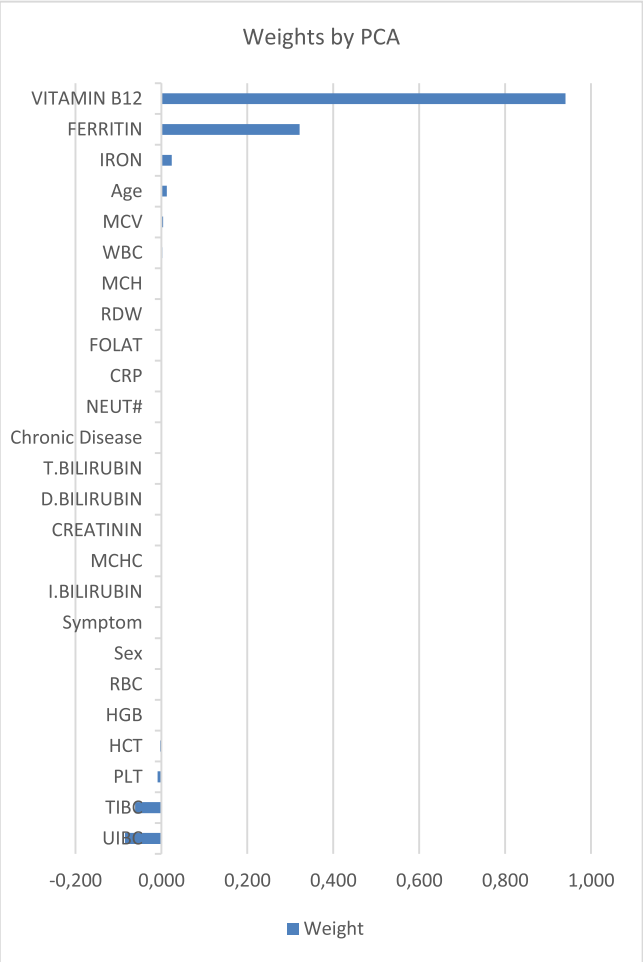

Fig. 9. Attribute weights by principal component analysis.

```
===Attribute Selection on all input data ===
Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 178
        Merit of best subset found:    0.549
Attribute Subset Evaluator (supervised, Class (nominal): Diagnosis):
        CFS Subset Evaluator
        Including locally predictive attributes
Selected attributes: 1,9,10,12,17,22,25 : 7
                     Age
                     FERRITIN
                     FOLATE
                     HGB
                     MCV
                     T. BILIRUBIN
                     VITAMIN B-12
```

Fig. 10. Attribute selection by CFS.

ROC analysis is illustrated in Table 8. Accuracy, recall/sensitivity, specificity, precision/confidence, F1-score, and AUC (area under the curve) values are basically calculated as in Equations (9) –(14).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$Recall/Sensitivity = \frac{TP}{TP + FN} \tag{10}$$

$$Specificity = \frac{TN}{TN + FP} \tag{11}$$

$$Precision/Confidence = \frac{TP}{TP + FP} \tag{12}$$

$$F1 - score = \frac{2.P.R}{P + R} \tag{13}$$

$$AUC = \frac{TPR - TNR}{2} \tag{14}$$

In Equations (9)–(14):

TP (True Positive): Number of samples when the predicted value and the real value are positive.

TN (True Negative): Number of samples when the predicted value and the real value are negative.

FP (False Positive): Number of samples when the predicted value is positive, and the real value is negative.

FN (False Negative): Number of samples when the predicted value is negative, and the real value is positive.

P: Precision/ Confidence

R: Recall/ Sensitivity

TPR (True Positive Rate): Sensitivity.

TNR (True Negative Rate): Specificity.

The ROC curve is used to evaluate the equilibrium between accuracy and sensitivity. The area remaining below the ROC curve,

Table 8
Confusion Matrix for ROC analysis.

		Real	
		Positive	Negative
Predicted	Positive	TP (True Positive)	FP (False Positive)
	Negative	FN (False Negative)	TN (True Negative)

known as the area under the curve (AUC), is defined as the ROC score. The ROC curve is plotted depending on the changing classification threshold values of true positives as a function of false positives. A ROC score of "1″ signifies that the positives are separated from the negatives in an excellent way. A ROC score of "0" means that no positives are found [61]. ROC analysis is widely used for binary-classed problems, yet it is also suitable for multi-class problems. A two-class approximation is used for multi-class problems. One of these approaches is the "one-versus-one" and the other is the "one-versus-all". In the "one-vs-one" approach, each class is compared pairwise to each of the others. In the "one-vs-all" approach, for class "t", all other classes are marked as "not t" and compared with "t". [62] In the present study, multi-class ROC analysis is applied using the "one-vs-one" approach. The precision is the rate at which the predicted positive class is actually positive. The recall is the ratio of correctly predicted true positives. Because these two metrics are also important performance metrics, the F1 score is also calculated using these two metrics [63]. The following section gives the resulting confusion matrices and ROC curves for all the methods used in this study (Figs. 10 –23).

4. Results

In this study, four machine learning methods are tested with original patient data to diagnose the 12 types of anemia seen in Table 2 in the same way as an experienced medical consultant would do. These are the most common types of anemia in Düzce Province where the data were collected.

The performance metrics in Table 9 are obtained when the original data are classified by the ANNs, various types of SVM, Naïve Bayes, and decision tree models. For this study, boosted and bagged decision tree models are used in addition to ANNs, linear, quadratic, and cubic SVM models, and a Naïve Bayes model.

As seen in Table 9, the highest accuracy rate is achieved by the boosted decision trees. In the AdaBoost method used in this study, the weights of misclassified samples of the decision tree are increased and the weights of correctly classified samples are reduced at each iteration. In the subsequent iterations, the updated weights are used. Thus, at each iteration, the algorithm is concentrated on misclassified cases. The ROC curves and confusion matrices of the most successful methods obtained with all datasets are shown in Figs. 11-24.

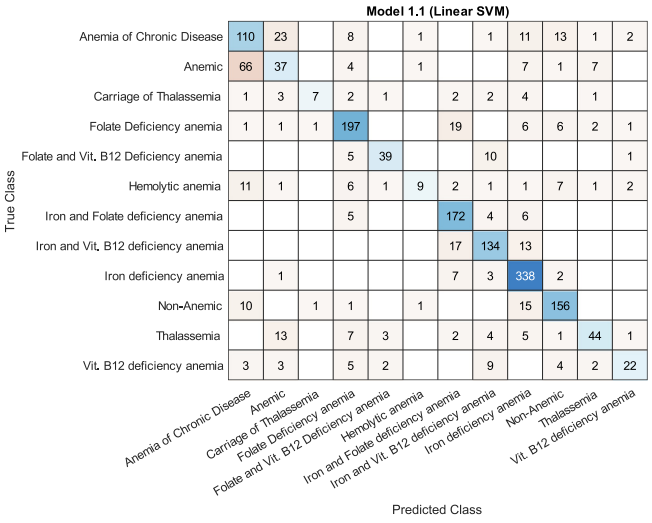

Fig. 11. Confusion matrix of the linear SVM with CFS-7 dataset.

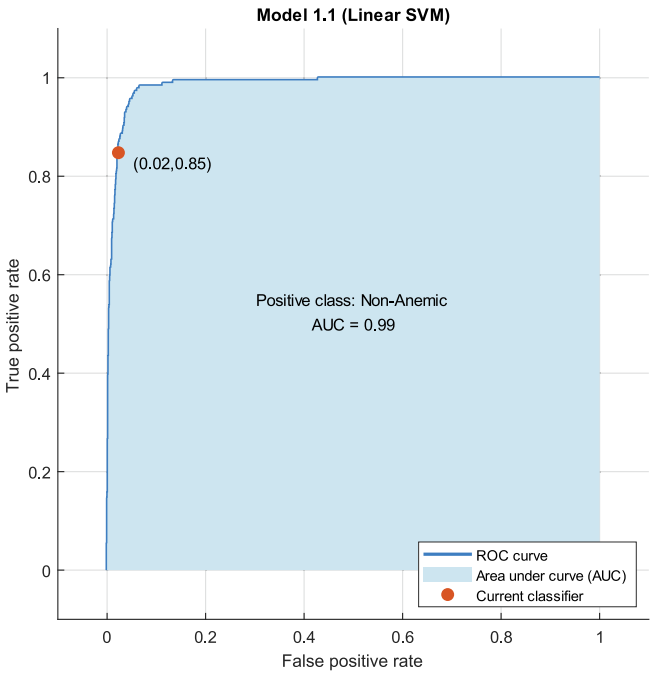

Fig. 12. ROC curve of the linear SVM with CFS-7 dataset.

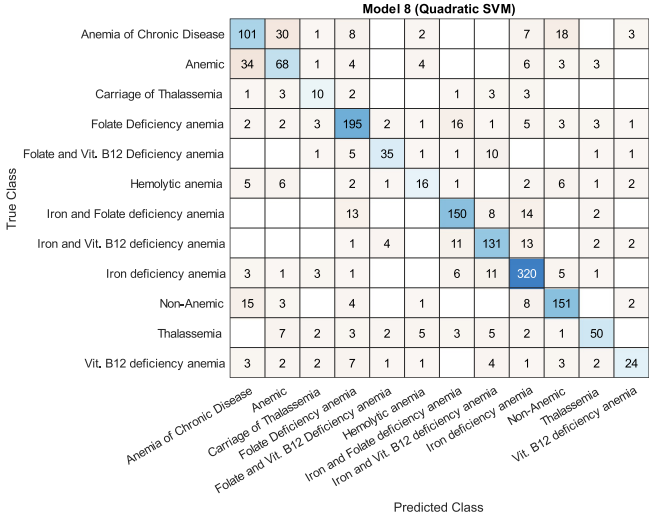

Fig. 13. Confusion matrix of the quadratic SVM with CFS-7 dataset.

Tables 10 – 19 show results when the methods are run on the datasets obtained by the feature selection process performed using the four different methods.

When the feature selection is completed in the dataset, it is obvious that the success of almost all methods had increased. Analysis of the performance for the 7-featured CFS (Correlation-based feature subset selection) dataset (Table 10) indicated that the success of the boosted trees and neural networks had decreased according to the original data. The most successful method is the bagged trees.

Tables 11 – 13 show results when the methods are run on the datasets obtained by the feature selection process performed using the information gain method.

Analysis of the performance for the 3-featured information gain dataset (Table 11) indicated that the success had decreased compared to the 7-featured CFS dataset (Table 10). However, more

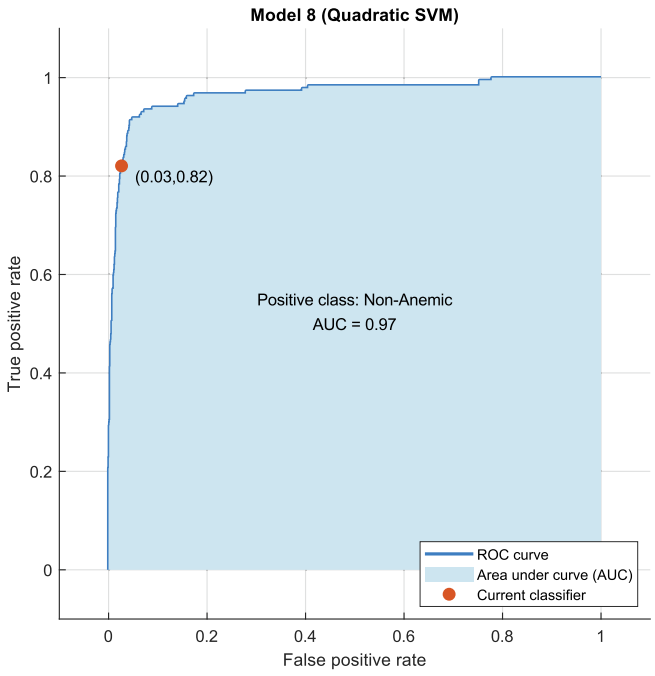

Fig. 14. ROC curve of the quadratic SVM with CFS-7 dataset.

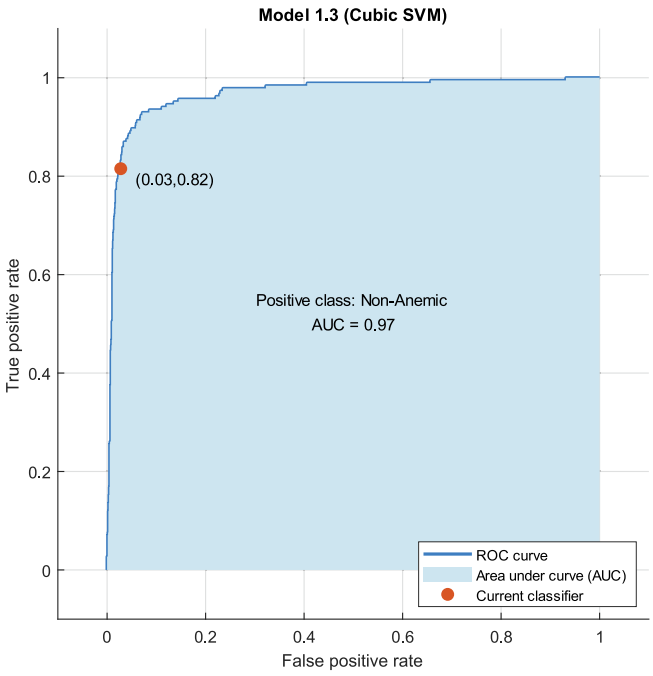

Fig. 16. ROC curve of the cubic SVM with info. gain-5 dataset.

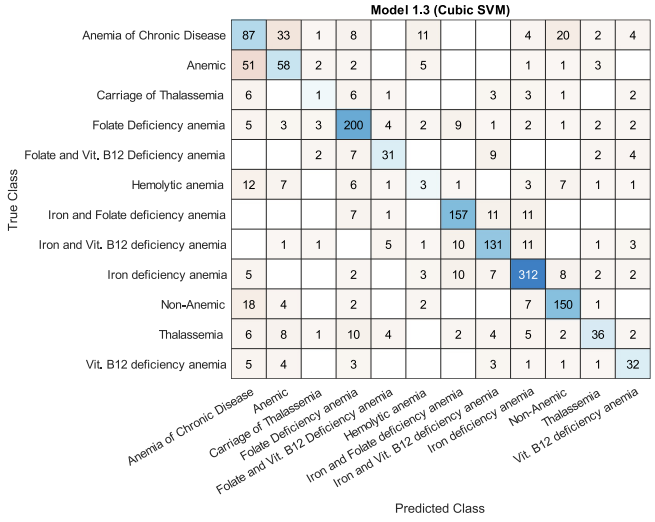

Fig. 15. Confusion matrix of the Cubic SVM with Info. Gain-5 dataset.

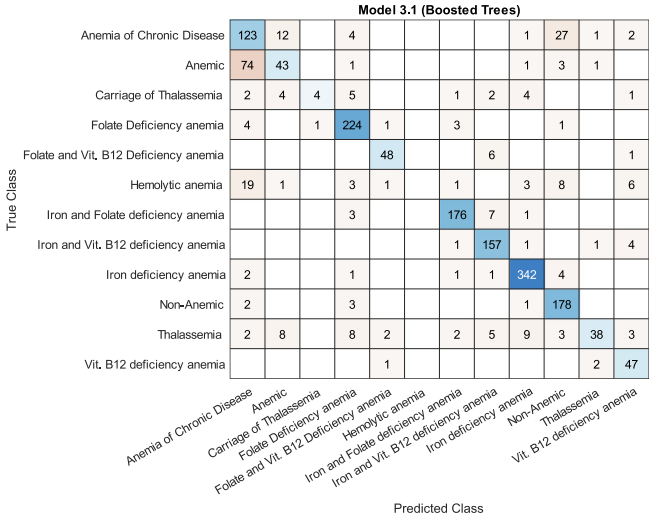

Fig. 17. Confusion matrix of the boosted tree with PCA-16 dataset.

successful results are obtained than with the original dataset. Although there is a decline in the success of the ANN and decision tree-based methods, the success of the Naïve Bayes and SVM methods increased.

Analysis of the performance for the 5-featured information gain dataset (Table 12) showed that the success had increased compared to the previous 3-featured information gain dataset (Table 11). This increase is noticeable for all methods. Although the attributes in the 3-featured dataset are the highest information gain attributes, using only these attributes did not increase the success of the methods.

Analysis of the performance table for the 13-featured information gain dataset (Table 13) demonstrated that the success is slightly lower than with the previous datasets. In the bagged and boosted decision tree methods, the increase in success stands out. With the increase in the number of features, a slight decrease is observed in the success of the SVM and Naïve Bayes methods.

Tables 14–16 show results when the methods are run on the datasets obtained by the feature selection process performed using the information gain ratio method.

Analysis of the performance for the information gain ratio-7 dataset indicated that the success is close to that of the previous datasets.

Analysis of the performance for the information gain ratio-11 dataset revealed no remarkable difference in the success of the methods compared with their success in the previous datasets.

Analysis of the performance for the information gain ratio-16 dataset again indicated that no remarkable increase had occurred in the success rate of the methods. However, in general, the decrease in the number of features led to a decline in the success rate of the ANN and DT-based methods as well.

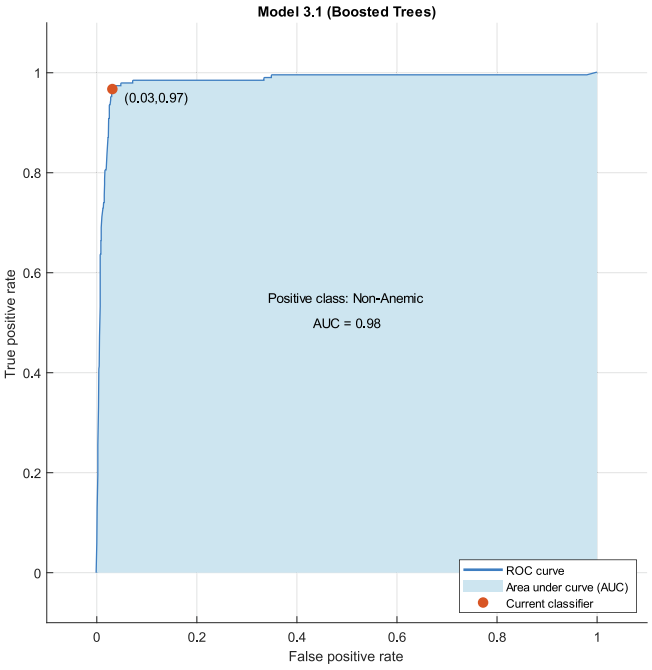

Fig. 18. ROC curve of the boosted tree with PCA-16 dataset.

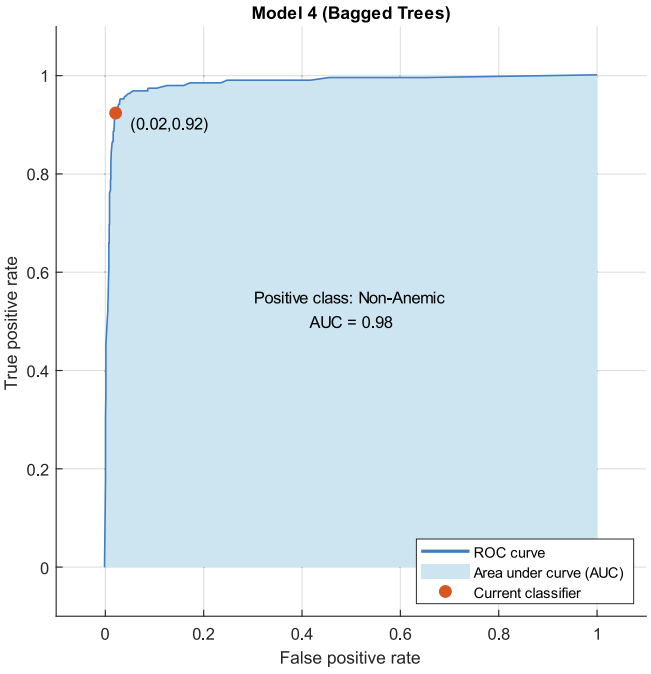

Fig. 20. ROC curve of the bagged tree with CFS-7 dataset.

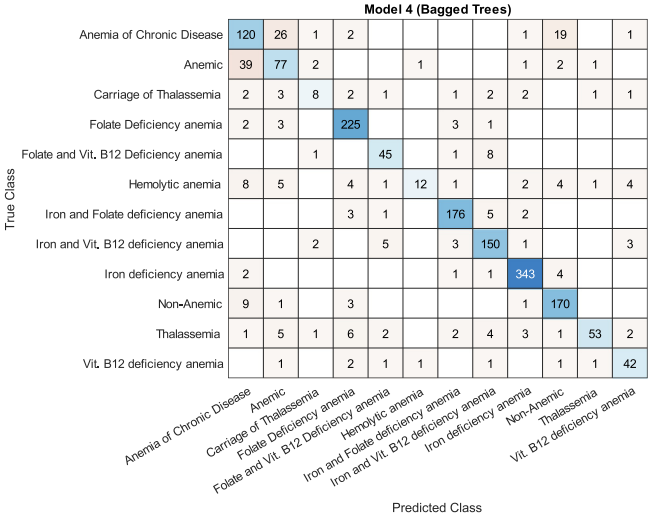

Fig. 19. Confusion matrix of the bagged tree with CFS-7 dataset.

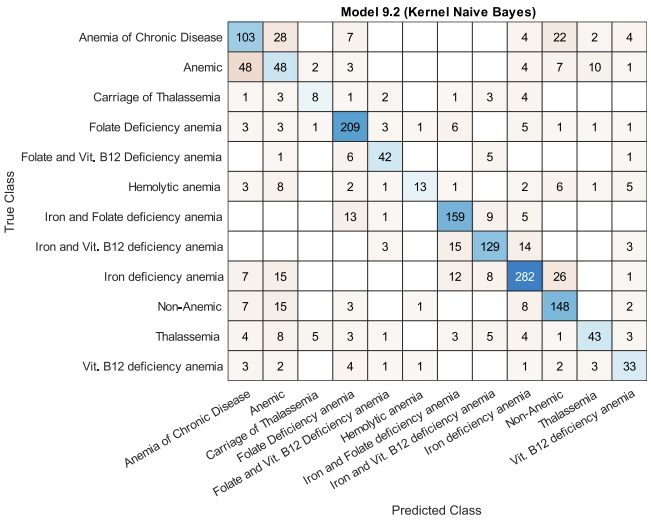

Fig. 21. Confusion matrix of the naïve Bayes with CFS-7 dataset.

Analysis of the performance table for the PCA-16 dataset (Table 17) showed a success rate similar to the previous datasets. The results obtained with this 16-featured dataset are more successful than the results obtained with the original 25-featured dataset.

Tables 18 and 19 present the datasets and models that yielded the most successful performance. Linear and Quadratic SVM, Bagged Tree, and Naïve Bayes methods demonstrated the highest success with the CFS-7 dataset. Cubic SVM showed the most success with the information gain-5 dataset. The Boosted Tree exhibited the most success with the PCA-16 dataset. The ANN method is the most successful with the original dataset, which is a notable finding.

The confusion matrices and ROC curves of the most successful methods used in the present study are shown in Figs. 11 - 24.

Figs. 11 and 12 show the most successful results of the linear SVM methods confusion matrix, and the ROC curve obtained with the CFS-7 dataset.

As seen in Figs. 11 and 12, linear SVM predicted the anemia types at a 76.1% accuracy rate. The highest success rate is achieved in the diagnosis of iron deficiency anemia (96.1%), and the lowest in the diagnosis of hemolytic anemia (21.4%).

Figs. 13 and 14 show the most successful results of the quadratic SVM methods confusion matrix, and the ROC curve, obtained with the CFS-7 dataset.

As seen in Figs. 13 and 14, quadratic SVM predicted anemia types at an accuracy rate of 75.8%. The highest success rate is in the diagnosis of patients with iron deficiency anemia (94.1%), whereas the lowest is in the diagnosis of hemolytic anemia (42.8%).

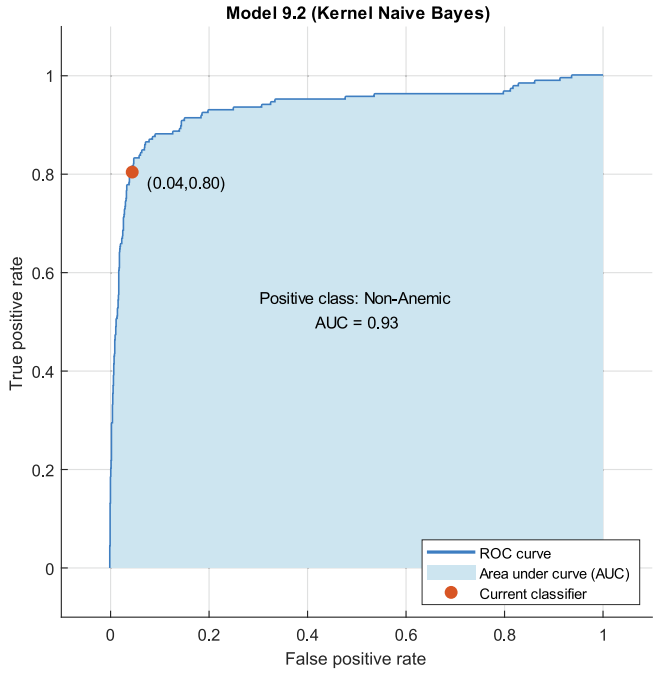

Fig. 22. ROC curve of the Naïve Bayes with CFS-7 dataset.

Figs. 15 and 16 show the most successful results of the cubic SVM methods confusion matrix, and the ROC curve, obtained with the InfoGain-5 dataset.

Figs. 15 and 16 show that cubic SVM predicted anemia types at a rate of 72% accuracy. The highest diagnosis success rate is achieved in patients with iron deficiency anemia (88.9%), while the lowest is in the diagnosis of thalassemia trait (carrier) patients (4.3%).

Analysis of the results showed that the highest prediction success rate using the SVM methods is obtained when the kernel function is selected as linear (76.1%). When the kernel function is selected as quadratic, the success rate (75.8%) is similar to the linear rate; however, the lowest prediction success rate is obtained when the kernel function is selected as cubic (72%). The disease classified at the highest rate using SVM models is iron deficiency anemia, whereas those classified at the lowest rate are hemolytic anemia and thalassemia trait.

Figs. 17 and 18 show the most successful results of the boosted decision tree methods confusion matrix, and the ROC curve, obtained with the PCA-16 dataset.

As seen in Figs. 17 and 18, the boosted decision tree predicted anemia types at an 83.0% accuracy rate. The highest success rate (97.4%) is achieved in the diagnosis of patients with iron deficiency anemia. The lowest success rate is seen in the diagnosis of hemolytic anemia (0%). This may be attrıbuted to the number of sam-

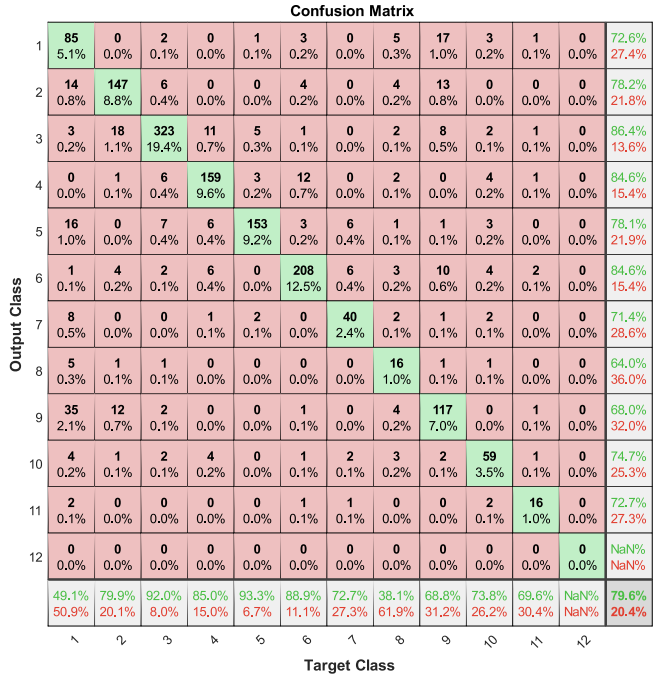

Fig. 23. Confusion matrix of the neural network with the original dataset.

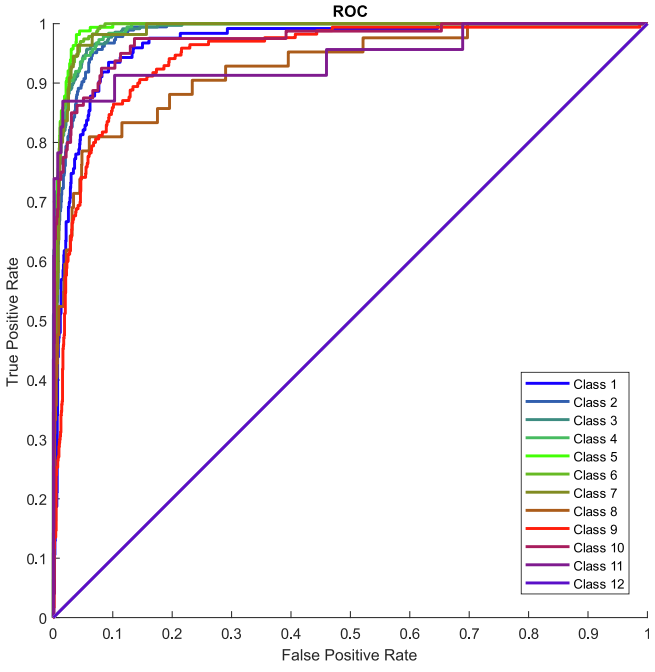

Fig. 24. ROC curve of the neural network with the original dataset.

Table 9
Performance of models for the original dataset including 25 features.

	Method	Accuracy	AUC	Precision	Recall	F1-score	error
Original Dataset (25 attributes)	Linear SVM	68.1	0.97	70.1	77.7	0.737	31.93
	Quadratic SVM	69.3	0.97	73.9	75.1	0.748	30.72
	Cubic SVM	64.3	0.96	72.8	70.5	0.716	35.71
	Boosted Trees	83.2	0.98	96.7	78.1	0.864	16.77
	Bagged Trees	80.6	0.98	92.9	79.2	0.855	19.42
	Naïve Bayes	59.8	0.88	76.6	52.8	0.625	40.22
	ANN	79.6	–	78.2	79.9	0.790	20.4

Table 10
Performance of models for the 7-featured CFS dataset.

	Method	Accuracy	AUC	Precision	Recall	F1-score	error
CFS Dataset(7 attributes)	Linear SVM	76.1	0.98	83.2	82.7	0.829	23.93
	Quadratic SVM	75.8	0.96	82.1	79.9	0.809	24.23
	Cubic SVM	71.8	0.95	78.3	74.6	0.764	28.20
	Boosted Trees	81.9	0.98	96.7	77.7	0.861	18.09
	Bagged Trees	85.6	0.98	90.8	82.3	0.863	14.37
	Naïve Bayes	74	0.93	81.5	66.7	0.733	26.03
	ANN	76.4	–	74.5	85.9	0.798	23.6

ples. This disease and thalassemia trait are among the lowest number of samples in the dataset

Figs. 19 and 20 show the most successful results of the bagged decision tree methods confusion matrix, and the ROC curve, obtained with the CFS-7 dataset.

As seen in Figs. 19 and 20, the bagged decision tree predicted anemia types at an 85.6% accuracy rate. The highest success rate is achieved in the diagnosis of iron deficiency anemia (97.7%), whereas the lowest is 35.7% for hemolytic anemia. The reason for this may have been, as mentioned above, the number of samples.

Figs. 21 and 22 show the most successful results of the Naïve Bayes methods confusion matrix, and the ROC curve, obtained with the CFS-7 dataset.

Figs. 21 and 22 show that Naïve Bayes predicted anemia types at an accuracy rate of 74%. The highest success rate is obtained in the diagnosis of folate deficiency anemia (87.6%), while the lowest is in the diagnosis of hemolytic anemia (26.1%).

Table 11
Performance of models for the 3-featured info-gain dataset.

	Method	Accuracy	AUC	Precision	Recall	F1-score	Error
Info. Gain Dataset(3 attributes)	Linear SVM	71.1	0.9	88.6	37.1	0.523	28.86
	Quadratic SVM	71.1	0.89	88.6	38.6	0.537	28.92
	Cubic SVM	66.9	0.86	22.3	34.7	0.271	33.07
	Boosted Trees	71.6	0.9	83.7	36.4	0.507	28.44
	Bagged Trees	71.6	0.88	64.7	37.8	0.477	28.38
	Naïve Bayes	66.5	0.75	49.5	39.2	0.437	33.49
	ANN	63.8	–	36.5	67.9	0.475	36.2

Table 12
Performance of models for the 5-featured info-gain dataset.

	Method	Accuracy	AUC	Precision	Recall	F1-score	Error
Info. Gain Dataset(5 attributes)	Linear SVM	75.5	0.98	86.4	78.7	0.823	24.53
	Quadratic SVM	74.6	0.97	83.7	80.2	0.819	25.37
	Cubic SVM	72	0.97	81.5	78.5	0.799	27.96
	Boosted Trees	79.9	0.98	96.7	78.4	0.865	20.14
	Bagged Trees	81.2	0.97	90.2	83.4	0.866	18.82
	Naïve Bayes	70.7	0.94	81.5	68.5	0.744	29.34
	ANN	72.9	–	78.6	82.1	0.803	27.1

Table 13
Performance of models for the 13-featured info-gain dataset.

	Method	Accuracy	AUC	Precision	Recall	F1-score	Error
Info. Gain Dataset(13 attributes)	Linear SVM	72.5	0.98	76.6	80.6	0.785	27.48
	Quadratic SVM	73.6	0.97	82.1	80.3	0.812	26.39
	Cubic SVM	68.7	0.96	78.8	71.8	0.751	31.26
	Boosted Trees	82.2	0.98	97.3	78.2	0.867	17.79
	Bagged Trees	83.8	0.98	91.8	80.5	0.857	16.23
	Naïve Bayes	65.1	0.91	79.3	58.9	0.675	34.87
	ANN	71.1	–	78.5	69.6	0.738	28.9

Table 14
Performance of models for the 7-featured info-gain-ratio dataset.

	Method	Accuracy	AUC	Precision	Recall	F1-score	Error
Info. Gain Ratio Dataset (7 attributes)	Linear SVM	72.1	0.94	72.3	60.7	0.659	27.90
	Quadratic SVM	71.3	0.91	66.3	55.7	0.605	28.74
	Cubic SVM	67.6	0.88	59.2	53.4	0.561	32.35
	Boosted Trees	77.6	0.94	78.3	64.6	0.707	22.36
	Bagged Trees	78.7	0.95	71.2	61.5	0.659	21.34
	Naïve Bayes	65.8	0.81	54.3	46.3	0.499	34.21
	ANN	65.1	–	48.1	28.3	0.356	34.9

Table 15
Performance of models for the 11-featured info-gain-ratio dataset.

	Method	Accuracy	AUC	Precision	Recall	F1-score	Error
Info. Gain Ratio Dataset (11 attributes)	Linear SVM	74	0.98	75.5	79	0.772	26.03
	Quadratic SVM	72.5	0.96	76.1	76.5	0.762	27.54
	Cubic SVM	67.6	0.95	73.4	69.2	0.712	32.41
	Boosted Trees	80	0.97	91.3	68.9	0.785	20.02
	Bagged Trees	80.8	0.97	83.7	74.4	0.787	19.24
	Naïve Bayes	65.1	0.87	66.3	58.4	0.620	34.87
	ANN	76.2	–	71.4	85.3	0.777	23.8

Table 16
Performance of models for the 16-featured info-gain-ratio dataset.

	Method	Accuracy	AUC	Precision	Recall	F1-score	Error
Info. Gain Ratio Dataset (16 attributes)	Linear SVM	71	0.98	75	80.2	0.775	29.04
	Quadratic SVM	70	0.97	79.3	78.5	0.788	30.01
	Cubic SVM	65.1	0.96	73.4	70.3	0.718	34.87
	Boosted Trees	82	0.98	96.7	78.1	0.864	18.03
	Bagged Trees	84.2	0.98	94	79.4	0.860	15.75
	Naïve Bayes	61.4	0.88	72.3	53.8	0.616	38.60
	ANN	74.7	–	74	77.2	0.756	25.3

Table 17
Performance of models for the 16-featured PCA dataset.

	Method	Accuracy	AUC	Precision	Recall	F1-score	Error
PCA Dataset(16 attributes)	Linear SVM	71.1	0.97	76.6	80.6	0.785	28.86
	Quadratic SVM	70.8	0.97	76.1	76.9	0.764	29.22
	Cubic SVM	66.3	0.96	76.6	73.8	0.751	33.67
	Boosted Trees	83	0.98	96.7	79.5	0.872	17.01
	Bagged Trees	85.4	0.98	91.3	81.6	0.861	14.61
	Naïve Bayes	59	0.9	80.4	54.8	0.651	41.01
	ANN	76.7	–	78.3	84.2	0.811	23.3

Table 18
Performance of datasets according to models.

Method	Dataset	Accuracy
Linear SVM	CFS -7	76.1%
Quadratic SVM	CFS -7	75.8%
Cubic SVM	Info. Gain -5	72.0%
Boosted Trees	PCA -16	83.0%
Bagged Trees	CFS -7	85.6%
Naïve Bayes	CFS -7	74.0%
ANN	Original -25	79.6%

Figs. 23 and 24 show the most successful results of the ANN methods confusion matrix, and the ROC curve, obtained with the original dataset.

Figs. 23 and 24 show that the ANN predicted anemia types at a 79.6% accuracy rate. The highest success rate is obtained in the diagnosis of iron deficiency anemia (86.4%), and the lowest in the diagnosis of hemolytic anemia (64.0%). Due to the structure of the ANNs, it is significant that the success rate increased in the diagnosis of iron deficiency anemia, with the highest number of samples, while it decreased in the diagnosis of hemolytic anemia, one of the diseases with the fewest number of samples.

According to these results for the present study, the ANN and ensemble decision tree methods produced more meaningful results than the Naïve Bayes and SVM methods, and the distribution of the prediction rates of each class is more balanced. The results also show that the most successfully classified diseases are iron deficiency anemia and folate deficiency anemia and those classified at the lowest success rate are hemolytic anemia and thalassemia trait. One of the reasons for this may have been the numbers of samples. The class with the largest number of samples in the dataset is iron deficiency anemia with 351 samples, followed by 234 samples for the folate deficiency anemia class. The class with the fewest number of samples is thalassemia trait, with only 23 samples, followed by 42 samples in the dataset for hemolytic anemia.

Table 19
Performance of models according to datasets.

Dataset	SVM	Accuracy	Ensemble Trees	Accuracy	Naïve Bayes Accuracy	Neural Network Accuracy
Original −25	Quadratic	69.3%	Boosted	83.2%	59.8%	79.6%
CFS −7	Linear	76.1%	Bagged	85.6%	74.0%	76.4%
Info. Gain −3	Linear/Quadratic	71.1%	Boosted/Bagged	71.6%	66.5%	63.8%
Info. Gain −5	Linear	75.5%	Bagged	81.2%	70.7%	72.9%
Info. Gain −13	Quadratic	73.6%	Bagged	83.8%	65.1%	71.1%
Info. Gain Ratio −7	Linear	72.1%	Bagged	78.7%	65.8%	65.1%
Info. Gain Ratio −11	Linear	74.0%	Bagged	80.8%	65.1%	76.2%
Info. Gain Ratio −16	Linear	71.0%	Bagged	84.2%	61.4%	74.7%
PCA −16	Linear	71.1%	Bagged	85.4%	59.0%	76.7%

When all the results are analyzed, they indicated that the decision tree-based ensemble methods had produced more successful results with these data than the ANN, naïve Bayes, or SVM methods, and that the estimation rates are distributed more evenly. Performing feature selection in the dataset is an important factor that increased success. Another notable point in these findings is that the most successful result in the ANN method was obtained using the original dataset. In the future, we are planning to conduct new studies on this subject.

5. Discussion

In this study, 12 different types of anemia described in the WHO International Classification of Disease (ICD) Codes are diagnosed. In addition, the study included 25 different attributes used by an experienced medical specialist to diagnose the anemias. The data used in this study are completely original and included age, sex, chronic diseases and symptoms as well as blood parameters. New datasets are also created using known feature selection methods such as Information gain, information gain ratio, PCA, and CFS. Valuable results are obtained in the experiments using these datasets.

The methods and results of the studies mentioned in related works are examined and a novel study is conducted for the new dataset. Although the studies reviewed may have overlapping profiles, the objective is not to make a performance comparison since the data are not used with the same approach. In the present study, all the data used are taken from the files of diseased individuals. Therefore, as no data are taken from healthy individuals, the term ‘‘non-anemic” is considered suitable for patients having other blood disorders that did not belong to one of the anemia types. When the study is expanded, other hematological disorders may be included in the estimation of diseases other than anemia. In order to treat a patient, first, it is necessary to determine the disease. Various methods have been developed to help experts in making this diagnosis. In the present study, four of the most familiar artificial learning methods (artificial neural networks, Support Vector Machines, naïve Bayes, and decision trees) are applied to a completely original dataset and the results are discussed.

In this study, the Ethics Committee permission is required to use real patient data. In order to obtain the approval of the Ethics Committee, every detail had to be determined. Therefore, all the attributes of the data had to be declared to the Committee. As there is no need for any pre-processing other than digitizing some data, no intervention is required. Because there is no provision for changing the content and quality of the data, in this study, a completely original dataset is used in which no numerical intervention had been made. Moreover, the use of real patient data made the relevance of this study significant. After obtaining the dataset, at the next stage, the data are classified using various common methods found in the literature that had been successful and had yielded significant results. During classification, we used the 10-fold cross-validation method, in which the dataset is divided into 10 different subsets. When one group is a testing set, the remaining nine groups served in turn as a training set. In this way, all the combinations are tested so that each of the 10 subsets had been a testing set once, and a performance value is found by taking the average of each result. The ROC analysis, precision, recall, F-score, AUC, and error metrics are used to evaluate performance. The outcome of the study also provided decision-making support, helping doctors and medical students in the diagnosis of 12 different types of anemia. The limitation of this study is that the entire dataset consisted of data from individuals with hematological diseases. For this reason, the data referred to as ‘‘non-anemic” are based on individuals with other diseases. At the same time, anemia is also a symptom that can accompany other diseases. Despite this limitation, it is obvious that the classifier performances we obtained are acceptable. We believe that this study makes a positive contribution to the literature.

6. Conclusion

Anemia is a very common disease affecting the quality of life and with the appropriate treatment, the standard of living of the patient will improve. It is obvious that the first step in the treatment is the correct diagnosis. In this study, the 12 types of anemia most commonly encountered in Düzce Province are classified by four different machine learning methods: ANNs, SVMs, naïve Bayes, and ensemble decision trees. The method with the highest success rate is the bagged decision tree method. Feature selection is performed using information gain, information gain ratio, principal component analysis (PCA), and correlation-based feature subset selection (CFS) methods. As a result of the feature selection process, nine different datasets are obtained. Thus, our dataset, which initially consisted of 25 attributes, is expanded to different datasets of 16, 13, 11, 7, 5, and 3 attributes. Here, our aim is to reach the doctor’s decisions with the data used by the doctor. Therefore, the original dataset is included in the datasets as well. All methods are also run on the new datasets and we compared the results of them all. The evaluation is carried out according to the accuracy, confusion matrix, ROC curves, classification error, precision, recall, and F-score metrics.

Finally, in order to improve the success rate, future studies should focus on combining different methods and on developing new hybrid methods. Moreover, by expanding the dataset, hematological diseases other than anemia could also be included in these studies.

Ethical Approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The data used in this study were taken from the patient files of the Düzce University Research and Application Hospital, Turkey, with the approval of the Düzce University Ethics Committee (Decision number 2014/82). Data were obtained without examining information on the identity of the patients.

Funding

This work was supported by the Research Fund of Sakarya University, Turkey, under Project Number: 2015–50-02–010. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] T. Vos et al., Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010, The Lancet 380 (9859) (2012) 2163–2196, https://doi.org/10.1016/S0140-6736(12)61729-2.

[2] WHO, ‘‘Haemoglobin concentrations for the diagnosis of anemia and assessment of severity, ‘‘Vitamin and Mineral Nutrition Information System, World Health Organization, Geneva, 2011.

[3] N.J. Kiassebaum et al., A systematic analysis of global anemia burden from 1990 to 2010, Blood 123 (2014) 615–624, https://doi.org/10.1182/blood-2013-06-508325.
[4] C. Fan, P. Chang, J. Lin, J.C. Hsieh, A hybrid model combining 4 case-based reasoning and fuzzy decision tree for medical data classification, Appl Soft Comput 11 (2011) 632–644.
[5] G. Orrù, W. Pettersson-Yeo, A.F. Marquand, G. Sartori, A. Mechelli, Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review, Neurosci Biobehav R 36 (2012) 1140–1152.
[6] J.R. Beck, J.R. Bell, F. Hirai, J.J. Simmons, and H.C. Jr. Lyon, "Computer-Based Exercises in Cardiac Diagnosis (PlanAlyzer)," Proc. Annu. Symp. Computer Applications in Medical Care 1988 Nov 9, pp. 403-408, PMCID: PMC2245328, 1988.
[7] J. Beck, J. ÓDonnell, F. Hirai, J. Simmons, J. Healy, H. Lyon, Computer-Based Exercises in Anemia Diagnosis (PlanAlyzer), Methods Inf Med 28 (04) (1989) 364–369, https://doi.org/10.1055/s-0038-1636782.
[8] H. C. Jr. Lyon, J.R. Bell, J.F. O'Donnell, F. Hirai, J.C. Healy, J.R. Beck, "The PlanAlyzer Cases for Teaching Clinical Reasoning: A Demonstration of the Cases, Discussion of the Research & Development Process, Lessons Learned and Strategies for Introducing Computer-Based Programs into Medical School Courses as a Vehicle for Curriculum Reform", Proc. Annu. Symp. Computer Applications in Medical Care 1993, PMCID: PMC2850769, 1993.
[9] M. Lipkin, Correlation of Data with a Digital Computer in the Differential Diagnosis of Hematological Diseases, IRE Trans. Med. Electron. ME-7 (4) (1960) 243–246, https://doi.org/10.1109/IRET-ME.1960.5008068.
[10] R.L. Engle, B.J. Flehinger, S. Allen, R. Friedman, M. Lipkin, B.J. Davis, L.L. Leveridge, HEME: A Computer Aid To Diagnosis Of Hematologic Disease, Bull. N. Y. Acad. Med. 52 (5) (1976) 584–600.
[11] I. Azarkhish, M.R. Raoufy, S. Gharibzadeh, Artificial Intelligence Models for Predicting Iron Deficiency Anemia and Iron Serum Level Based on Accessible Laboratory Data, J Med Syst 36 (3) (2012) 2057–2061, https://doi.org/10.1007/s10916-011-9668-3.
[12] Z. Yılmaz, M.R. Bozkurt, Determination of Women Iron Deficiency Anemia Using Neural Networks, J Med Syst 36 (5) (2012) 2941–2945, https://doi.org/10.1007/s10916-011-9772-4.
[13] A. Yılmaz M. Dağlı N. Allahverdi A Fuzzy Expert System Design for Iron Deficiency Anemia 2013 Baku, Azerbaijan pp. 1 – 4. 10.1109/ICAICT.2013.6722707
[14] W. Reinisch, M. Staun, S. Bhandari, M. Munoz, State of the iron: How to diagnose and efficiently treat iron deficiency anemia in inflammatory bowel disease, Journal of Crohn's and Colitis 2013 (7) (2013) 429–440.
[15] Ş. Doğan, I. Türkoğlu, Iron-Deficiency Anemia Detection from Hematology Parameters by Using Decision Trees, International Journal of Science & Technology 3 (1) (2008) 85–92.
[16] B.Ç. Yavuz T.K. Yıldız N. Yurtay Z. Pamuk Comparison Of K Nearest Neighbours And Regression Tree Classifiers Used With Clonal Selection Algorithm To Diagnose Haematological DiseasesComparison Of K Nearest Neighbours And Regression Tree Classifiers Used With Clonal Selection Algorithm To Diagnose Haematological Diseases AJIT-e 5 16 7 20 10.5824/1309-1581.2014.3.001.x
[17] S.A. Sanap M. Nagori V. Kshirsagar Classification of Anemia Using Data Mining Techniques, Presented in SEMCCO, Berlin Heidelberg Part II 2011 113 12
[18] N. Amin, A. Habib, Comparison of Different Classification Techniques Using WEKA for Hematological Data, Am J Eng Res 4 (3) (2015) 55–61. e-ISSN: 2320-0847.
[19] I. Yılmaz, Demir eksikliği anemisi ile beta talasemi minörün ayırıcı tanısında eritrosit indekslerin rolü Expertise Thesis, Dept. of Internal Medicine, Faculty of Medicine, Pamukkale University, Denizli, Turkey, 2010.
[20] E. Urrechaga, U. Aguirre, S. Izquierdo, Differential Diagnosis of Microcytic Anemia, Anemia (2013), https://doi.org/10.1155/2013/457834.
[21] M.K. Jamei, K.M. Talarposhti, Discrimination between Iron Deficiency Anaemia (IDA) and β - Thalassemia Trait (β-TT) Based on Pattern-Based Input Selection Artificial Neural Network (PBIS- ANN), J Adv Comp Res 7 (4) (2016) 55–66. pISSN: 2345-606x, eISSN: 2345-6078.
[22] R. Kishore, K.P. Rao, G.R.S. Murthy, Performance Evaluation of Entropy and Gini using Threaded and Non-Threaded ID3 on Anaemia Dataset, Presented at Fifth International Conference on Communication Systems and Network Technologies, 2015.
[23] M. F. Shaik and M. Subashini, Anemia Diagnosis by Fuzzy Logic Using LabVIEW, Presented at IEEE International Conference on Intelligent Computing and Control (I2C2), 2017. DOI: 10.1109/I2C2.2017.8321790
[24] P.T. Dalvi, N. Vernekar, Anemia Detection using Ensemble Learning Techniques and Statistical Models, Presented at IEEE International Conference on Recent Trends in Electronics Information Communication Technology, May, 2016.
[25] S. Belginova I. Uvaliyeva A. Ismukhamedova Decision Support System for Diagnosing Anemia Presented at 4th International Conference on Computer and Technology Applications 2018 10.1109/CATA.2018.8398684
[26] G. Dimauro, D. Caivano, F. Girardi, A New Method and a Non-Invasive Device to Estimate Anemia Based on Digital Images of the Conjunctiva, IEEE Access 6 (2018) 46968–46975, https://doi.org/10.1109/ACCESS.2018.2867110.
[27] J. YEH, C. CHENG, Using hierarchical soft computing method to discriminate microcyte anemia, Expert Syst. Appl. 29 (3) (2005) 515–524, https://doi.org/10.1016/j.eswa.2005.04.012.
[28] N. Allahverdi, A. Tunali, H. Işık, H. Kahramanli, A Takagi-Sugeno type neuro-fuzzy network for determining child anemia, Expert Syst. Appl. 38 (2011) 7415–7418.
[29] M. Maity P. Sakar C. Chakraborty Computer Assisted Approach to Anemic Erythrocyte Classification Using Blood Pathological Information, Emerging Applications of Information Technology (EAIT) ISBN:978-1-4673-1827-3112, 2012. India 116 121
[30] D. Setsirichok, T. Piroonratana, W. Wongseree, T. Usavanarong, N. Paulkhaolarn, C. Kanjanakorn, M. Sirikong, C. Limwongse, N. Chaiyaratana, Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naive Bayes classifier and a multilayer perceptron for thalassaemia screening, Biomed. Signal Process. Control 7 (2012) 202–212.
[31] K. Meena, D.K. Tayal, V. Gupta, A. Fatima, Using classification techniques for statistical analysis of Anemia, Artif. Intell. Med. 94 (2019) 138–152, https://doi.org/10.1016/j.artmed.2019.02.005.
[32] G.N. Balaji, T.S. Subashini, N. Chidambaram, Detection and diagnosis of dilated cardiomyopathy and hypertrophic cardiomyopathy using image processing techniques, Engineering Science and Technology, an International Journal 19 (4) (2016) 1871–1880, https://doi.org/10.1016/j.jestch.2016.10.001.
[33] L. Shen, H. Chen, Z. Yu, W. Kang, B. Zhang, H. Li, B. Yang, D. Liu, Evolving support vector machines using fruit fly optimization for medical data classification, Knowl.-Based Syst. 96 (2016) 61–75, https://doi.org/10.1016/j.knosys.2016.01.002.
[34] M. Wang, H. Chen, B. Yang, X. Zhao, L. Hu, Z. Cai, H. Huang, C. Tong, Toward an optimal kernel extreme learning machine using a chaotic moth-flame optimization strategy with applications in medical diagnoses, Neurocomputing 267 (2017) 69–84, https://doi.org/10.1016/j.neucom.2017.04.060.
[35] M. Wang, H. Chen, Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis, Appl. Soft Comput. 88 (2020) 105946, https://doi.org/10.1016/j.asoc.2019.105946.
[36] C. Verma, V. Stoffova, Z. Illes, S. Tanwar, N. Kumar, IEEE Access, Vol.8, 2020
[37] C. Verma, V. Stoffova, Z. Illes, Prediction of residence country of student towards information, communication, and mobile technology for real-time: preliminary results, Procedia Computer Science (ICCIDS 167 (pp224–234) (2019) 2020.
[38] Bevilacqua V. et al. (2019) A Novel Deep Learning Approach in Haematology for Classification of Leucocytes. In: Esposito A., Faundez-Zanuy M., Morabito F., Pasero E. (eds) Quantifying and Processing Biomedical and Behavioral Signals. WIRN 2017 2017. Smart Innovation, Systems and Technologies, vol 103. Springer, Cham. https://doi.org/10.1007/978-3-319-95095-2_25V.
[39] V. Bevilacqua, A. Brunetti, A. Guerriero, G.F. Trotta, M. Telegrafo, M. Moschetta, A performance comparison between shallow and deeper neural networks supervised classification of tomosynthesis breast lesions images, Cognit. Syst. Res. 53 (2019) 3–19, https://doi.org/10.1016/j.cogsys.2018.04.011.
[40] R.S. Hillman, K.A. Ault, M. Leporrier, H.M. Rinder, Haematology in Clinical Practice. Part 1: Red Blood Cell Disorders, Chapter 2. Clinical Approach to Anemia, 5th ed., Güneş Medical Bookstore, Istanbul, Turkey, 2012.
[41] T. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. 27 (8) (2006) 861–874, https://doi.org/10.1016/j.patrec.2005.10.010.
[42] Guide to diagnosis and treatment of Erythrocyte Diseases and Hemoglobin disorders, Turkish Hematology Association, Version1-July2011. Available: www.thd.org.tr.
[43] R. Hoffman, Jr. E. J. Benz, L.E. Silberstein, H.E. Heslop, J.I. Weitz, J. Anastasi, Hematology: Basic Principles and Practice, 6th edition, ISBN: 978-1-4377-2928-3, Elsevier, 2013
[44] Turkish Hematology Association Erythrocyte Diseases and Hemoglobin Disorders Diagnosis and Treatment Guidelines, Version 1, Ankara, July 2011.
[45] Access Date 24.01.2014.
[46] Clinical Hematology, Ankara University, Faculty of Medicine, Department of Hematology, ISBN: 975-7226-35-1, Publications, ANTIP A.Ş, 1997.
[47] Hematology Laboratory Guide 1, II. Department: Complete Blood Count, Turkish Hematology Association, 2014
[48] International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10)-WHO Version for 2019, https://icd.who.int/browse10/2019/en#/III Access Date: 11.07.2020
[49] http://www.kanhastaliklari.org.tr/icerik.php?id=127&alt_id=196&tab=126 Access Date: 11.07.2020
[50] https://www.mayoclinic.org/diseases-conditions/iron-deficiency-anemia/symptoms-causes/syc-20355034 Access Date: 11.07.2020.
[51] https://www.mayoclinic.org/diseases-conditions/vitamin-deficiency-anemia/symptoms-causes/syc-20355025 Access Date: 11.07.2020
[52] http://www.thd.org.tr/thdData/userfiles/file/9_MSEK_09.pdf Access Date: 11.07.2020.
[53] http://thd.org.tr/thdData/userfiles/file/Ertitrosit-Tani-ve-tedavi-Kilavuzu-2019.pdf Access Date: 11.07.2020.
[54] E. Öztemel, Artificial Neural Networks, Papatya Press, Istanbul, Turkey, 2006.
[55] A.E. Hassanien, E.T. Al-Shammari, N. Ghali, Computational Intelligence Techniques in Bioinformatics, Comput Biol Chem 47 (2013) 37–47, https://doi.org/10.1016/j.compb iolche m. 2013.04.007.
[56] MATLAB User's Guide, MATLAB R2017a, Natick, MA, USA, The MathWorks Inc.
[57] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, D. Steinberg, Top 10 algorithms in data mining, Knowl Inf Syst 14 (1) (2008) 1–37, https://doi.org/10.1007/s10115-007-0114-2.
[58] C. E. Shannon A Mathematical Theory of Communication 27 4 1948 623 656 10.1002/j.1538-7305.1948.tb00917.x

[59] J.R. Quinlan, Induction of decision trees, Mach Learn 1 (1) (1986) 81–106, https://doi.org/10.1007/BF00116251.
[60] M.A. Hall, Correlation-based feature subset selection for machine learning, Doctor of Philosophy thesis in, the University of Waikato, Hamilton, New Zealand, 1999.
[61] D. J. Hand, R. J. Till, A simple generalization of the area under the ROC curve for multiple class classification problems, Machine Learning, Vol.45 pp:171-186, 2001.
[62] M.S. Wandishin S.J. Mullen Multiclass ROC analysis Weather and Forecasting 24 2008 10.1175 / 2008WAF2222119.1
[63] F. Ramzan, M.U.G. Khan, A. Rehmat, S. Iqbal, T. Saba, A. Rehman, Z. Mehmood, A Deep Learning Approach for Automated Diagnosis and Multi-Class Classification of Alzheimer's Disease Stages Using Resting-State fMRI and Residual Neural Networks, J Med Syst 44 (2) (2020), https://doi.org/10.1007/s10916-019-1475-2.