Neural Computing and Applications (2024) 36:5653–5672
https://doi.org/10.1007/s00521-023-09379-y

ORIGINAL ARTICLE

Classification of anemia using Harris hawks optimization method and multivariate adaptive regression spline

Nagihan Yagmur[1] · İdiris Dag[1] · Hasan Temurtas[2]

Received: 21 April 2023 / Accepted: 7 December 2023 / Published online: 10 January 2024

Abstract
Data mining methods are important for the diagnosis and prediction of diseases. Early and accurate diagnosis of patients is vital for their treatment. Various methods have been used in the literature to classify anemia. However, due to the different characteristics of patient datasets, changes in dataset sizes, different parameter numbers and features, and different numbers of patient records, algorithm performances vary according to datasets. In this study, the Harris hawks algorithm (HHA) and the multivariate adaptive regression spline (MARS) were used to classify anemia based on blood data of 1732 patients from the Kaggle database of patients with and without anemia. Six different algorithms were proposed to determine the parameters of the linear anemia approximation, namely multilinear form HHA, multilinear quadratic form HHA, multilinear exponential form HHA, first-order MARS model, second-order MARS model, and the best performing MARS model. The performance of the six proposed algorithms has been analyzed and found to be better than the previous studies in the literature.

Keywords Anemia prediction · Classification · Harris hawks optimization method · Multivariate adaptive regression spline

1 Introduction

Anemia is a global health problem affecting human health [1]. It particularly affects young children and pregnant women. The World Health Organization estimates that 42% of children under 5 years of age and 40% of pregnant women worldwide are anemic [2].

Anemia, which is expressed by a decrease in the number of red blood cells in the blood, occurs with a decrease in the level of hemoglobin in the blood with parameters such as sex, age, pregnancy, and nutrition. So anemia is defined as the hemoglobin value below the appropriate reference range. The measurement of the values related to the cells in the blood circulation is called the complete blood count (hemogram). The complete blood counting marks the blood values low or high according to the reference range.

Both diagnosis and treatment of the anemia is decided by doctors. In order to diagnose anemia more accurately, blood tests, radiological images, etc., must be observed by the doctor. The diseases produce a lot of medical data from which alternative solutions are produced such as to detect diseases at an early stage, to prescribe appropriate drugs to the patient, and not to extend the initial phase before reaching the critical phase. Consequently, disease determinations can be made for new patients according to the medical data obtained from the patients. This is very important for doctors to minimize the margin of error in the diagnosis they will make for the patient, and it is important for helping doctors to diagnose. Therefore, the evaluation of data records in health institutions is of great importance for patients and hospitals. However, these processes can be difficult and costly, especially in underdeveloped countries.

There are many studies on designing decision support systems for doctors for new patients by evaluating data records in hospitals with biomedical image processing,

✉ Nagihan Yagmur
nagihan.yagmur@dpu.edu.tr

İdiris Dag
idag@ogu.edu.tr

Hasan Temurtas
hasan.temurtas@dpu.edu.tr

[1] Mühendislik Mimarlık Fakültesi, Eskişehir Osmangazi Üniversitesi, Eskişehir, Turkey

[2] Mühendislik Fakültesi, Kütahya Dumlupınar Üniversitesi, Kütahya, Turkey

biomedical signal processing, biomedical digital data processing, etc. [3–6]. In image processing studies, medical images (magnetic resonance imaging (MRI), computerized tomography (CT) scans, etc.) have been analyzed and systems have been developed to help doctors make better treatment decisions [7–9]. Signal processing studies aim to develop systems that help doctors by analyzing and interpreting medical signals (electrocardiography (ECG), electroencephalography (EEG), etc.) [10, 11]. In studies on digital data processing, digital data (blood count, C-reactive protein (CRP) level, etc.) from patients are usually processed and systems have been developed to help doctors respond faster and more accurately for new patients. In addition to classical methods such as support vector machines, Naïve Bayes, regression, and k-nearest neighborhood, artificial intelligence-based methods such as artificial neural networks, deep learning, and random forest trees have started to be used in studies [12, 13].

Optimization methods have an important place in the solution of engineering problems. Modeling a problem has become an area where optimization methods are frequently used. Finding the model parameters that best represent the problem is a very important step for modeling the problem. For this reason, mathematical modeling is needed in areas such as data analysis, control system design, machine learning, etc. [14–16].

Engineering problems are faced with increasing levels of complexity day by day. Classical methods cannot be successful in the optimization of complex systems due to problems such as difficulty in solving high dimensional problems, local minima problems, the fact that many classical methods are designed for differentiable problems, etc. Therefore, the need for new optimization methods inspired by nature is increasing. These methods, which tend to perform better on complex problems, can deal with non-continuous problems and are less sensitive to local minima [17, 18]. Examples of nature-inspired algorithms frequently used in optimization are crow search optimization (CSO), chicken swarm algorithm (CSA), JAYA, ant colony, HHA, artificial bee colony (ABC), etc. [19, 20].

MARS method, which is another mathematical modeling method preferred for analyzing complex datasets, has been frequently used in prediction, analysis, classification, etc., studies [21]. When the studies in which the Mars method is applied are examined, it shows that this machine learning approach can help to create good prediction models for engineering datasets [22–24].

These methods may not perform well for every dataset due to different features in different datasets used in studies. Limits such as different parameter properties, number of parameters, and changes in the number of patient records in the dataset significantly affect the success of anemia disease classification methods. So, it is essential to develop new techniques because the properties of the studied datasets and the number of parameters or sizes may differ [13]. In machine learning, image, biomedical, robotics, natural language processing, and other fields, both classical and metaheuristic methods have been successfully used to classify data with different parameters and feature structures [25–27]. The methods have been developed from different perspectives such as linear, quadratic, and exponential in the literature, and the methods have been analyzed under various scenarios such as linear, quadratic, and exponential in order to model the relationships between the parameters in the datasets [28, 29]. The model weight values in the constructed scenarios were calculated by optimizing the proposed methods according to the objective function. Disease classification was made by generating the weight values with the lowest error.

For these reasons, new approaches and algorithms need to be developed to predict anemia. There are many data mining methods used in anemia diagnosis in the literature. These are: learning vector quantization neural network (LVQ), k-nearest neighbors (k-NN), multiple linear regression (MLR), logistic regression (LR), fuzzy logic (FL), artificial neural networks (ANN), etc. [30, 31].

In this study, 1732 blood data from the Kaggle database were analyzed using the Harris hawks algorithm, a nature-inspired evolutionary algorithm, and the MARS algorithm, a classical mathematical modelling method. The proposed methods are analyzed under 6 different scenarios: multilinear form HHA, multilinear quadratic form HHA, multiexponential form HHA, first-order MARS model, second-order MARS model, and MARS model to obtain the best degree and pruning coefficient. Thus, the pruning parameter and degree values, which have a significant effect on the performance of the MARS method in 3 models, enable the model to learn different relationships and reveal complex models, while in the other 3 models, the effects of the parameters on the classification success of the problem modelled in linear, quadratic, and exponential form were optimized by HHA method and the most appropriate weight values were obtained. To the best of our knowledge, no anemia classification study has been performed using the MARS method and parameter estimation method based on mathematical modelling with HHA.

2 Literature review

With the help of artificial neural networks and decision trees developed by genetic programming, an average of 90% performance was obtained as a result of the tests performed for the classification problem of thalassemia (Mediterranean anemia) disease [32]. In a 2008 study, a decision support system was designed to help physicians in

iron deficiency anemia [33]. Finally, Anemia (+) and Anemia (−) results were evaluated at the end of the procedure. The results of the decision support system completely coincided with the decisions of the doctors. Serum iron, serum iron-binding capacity, and ferritin were used as parameters in the study, and six different blood parameters, namely HGB, RBC, MCH, MCHC, WBC, and HCT, were used in our study. In a study conducted in 2011, anemia prediction and classification were analyzed using data mining techniques, J48 and sequential minimum optimization (SMO) classification methods were applied in Weka, and the C4.5 decision tree algorithm (CDTA) and support vector machine (SVM) were studied [34]. Another study designed a neuro-fuzzy network to determine the level of anemia in a child [35]. With this system, which was developed after statistical measurements, the root mean square of the errors was found to be 0.2743. In 2012, artificial neural networks and an adaptive neuro-fuzzy inference system (ANFIS) were developed to predict iron deficiency anemia based on four laboratory data of mean erythrocyte volume (MCV), mean cellular hemoglobin (MCH), mean cellular hemoglobin concentration (MCHC), and red blood cell count (RBC) [36]. In a study on iron deficiency anemia in women, feedforward networks (FFN), cascade forward networks (CFN), distributed delay networks (DDN), probabilistic neural network (PNN), and LVQ were used [37]. Another article presents the classification of blood characteristics with a CDTA, Bayesian classifier, and a multilayer perceptron (MP) [38]. With the study classified eighteen thalassemia anemia with high prevalence in Thailand, the best classification performance is obtained with the Naïve Bayes (NB) classifier and then with the multilayer perceptron. A study was carried out using machine learning algorithms in the detection of anemia [39]. In this study, ANN, SVM, and statistical model methods were applied in the diagnosis of iron deficiency. Some classification algorithms such as NB, MP, J48, and SMO were used by using WEKA data mining tool [40]. As a result, it was observed that the J48 decision tree algorithm (JDTA) had the best performance. The deep learning methodologies were used to increase the performance of white blood cell (WBC) identification systems. A new WBC recognition system has been proposed based on deep learning theory [41]. In a 2018 study, an easy-to-use and inexpensive device was developed to determine the anemia status in patients, preventing the patient from going to the laboratory frequently, allowing a large number of people to be screened for anemia [42]. It has been observed that there is a strong correlation between the information estimated by the device and the actual Hb values obtained by taking blood samples. The k-NN classification algorithm was used to assess the anemia status and gave good results. Thus, doctors avoid a significant number of blood tests [42]. In the study conducted in 2019, the effect of biochemistry values on iron deficiency anemia was investigated by k-NN, CDTA, and ANN methods, based on the blood values stated in the literature to be effective for iron deficiency anemia [43]. As a result, it has been seen that the highest-performance artificial neural network method is. In another study, machine learning algorithms, linear discriminant analysis (LDA), classification and regression trees (CART), SVM, randomized forest (RF), k-NN, and LR were used [44]. They found that the RF algorithm achieved the best classification accuracy. In another study conducted in 2019, a new machine learning method (HEAC—Hemoglobin Estimation and Anemia Classification) was proposed for anemia classification based on blood parameters and compared with other machine learning methods in the literature [45]. Since the symptoms of iron deficiency anemia and β thalassemia are similar in the study conducted in 2020, a decision support system was developed to ensure discrimination [31]. In the proposed system, LR, k-NN, SVM, extreme learning machine, and regularized extreme learning machine classification algorithms are used. As a result, in the study in which male and female patients were evaluated together, an accuracy of 96.30% for women and 94.37% for men was obtained. In a study conducted in 2021, a structure was proposed that will enable the recognition of anemia in clinical practice conditions [46]. ANN, SVM, NB, and ensemble decision tree methods were used as classification algorithms. In another study, two hybrid models using genetic algorithm (GA) and deep learning algorithms (DLA) of stacked autoencoder (SAE) and convolutional neural network (CNN) were proposed for the prediction of some types of anemia [13]. When the performances of the proposed algorithms were evaluated, the performance of the GA-CNN algorithm was found to be better. One study in 2022 used synthetic minority over-sampling technique SMOTE to improve the imbalance of the anemia dataset from India [47]. Then, with the help of the decision tree rule-based learning method, the rules for the detection of anemia were derived using the original and SMOTE dataset.

When the studies on anemia are examined, it is seen that the use of swarm-based optimization methods is quite low. This study aims to see the success of the HHA algorithm, one of these algorithms, also called metaheuristics, which uses the advantages of swarm intelligence to solve complex optimization problems that cannot be solved by analytical methods, to obtain weight coefficients that will emphasize the importance of the parameters in the dataset, and the success of the MARS method, a classical mathematical modeling method preferred for the analysis of complex datasets, in the classification problem with different degree and pruning parameters. Both methods are tested under three different scenarios to highlight their success by

modeling the relationships between dataset parameters. Both methods are tested under three different scenarios, and their success is highlighted by modelling the relationships between dataset parameters.

3 Material and method

3.1 Dataset and preprocessing

Blood data of 1732 patients from the Kaggle database were used in the study. The dataset consists of 351 patients with anemia and 1381 patients without anemia. As shown in Table 1, the study used 6 attributes and 2 classes, anemia (1)/healthy (0). The RBC value indicates the amount of red blood cells in the blood of each patient data, hemoglobin, HGB value indicates the amount of iron-rich protein stored in red blood cells, HCT value indicates the volumetric amount of blood in red blood cells, MCV value indicates the average cell volume, MCH value indicates the ratio of hemoglobin to red blood cells in a given volume, MCHC value indicates the average amount of hemoglobin in a single red blood cell, and 6 different blood components and their corresponding anemia outcome information.

3.2 Harris hawks optimization method

The Harris algorithm is an algorithm that works by imitating the hunting strategy of hawks in 2019 and is presented by Heidari as mathematically modeled [48]. When the literature is searched, it is seen that the Harris hawks method is used in many different areas [49–52]. However, the use of HHA in studies on disease prediction in the health sector has been limited [53–55]. Harris hawks move in packs with a leader at their head when hunting rabbits. First of all, they determine the location of the prey by making reconnaissance flights. They then move on to the hunting stage. This algorithm is population-based and consists of many stages.

Table 1 List of attributes in the dataset

Attribute name	Type	Min	Max	Avg
HGB	Numeric	2	13.100	4.484
RBC	Numeric	3.600	16.700	12.453
MCH	Numeric	15.700	52.110	38.003
WBC	Numeric	2.900	116	84.907
MCHC	Numeric	13.6	39.54	27.884
HCT	Numeric	22.9	43.7	32.713

3.2.1 Exploration phase

During the exploration phase, the Harris hawks wait and observe. This event continues in a loop. In each cycle, the hawk in the best position gives the best solution according to the position of the prey. While hawks wander, they make 2 different discoveries. These discoveries are given in Eq. 1. The value of q in the equation is a probabilistic value and indicates which discovery will be applied [48].

$$x(t+1) = \begin{cases} x_{\text{rand}}(t) - r_1 |x_{\text{rand}}(t) = 2r_2 x(t)|, q \geq .5 \\ (x_{\text{rabbit}}(t) - x_m(t)) - r_3(\text{LB} + r_4(\text{UB} - \text{LB})), q < .5 \end{cases} \quad (1)$$

In the equation, $x(t+1)$ is the vector indicating the position of the Harris hawk in each iteration, x_{rabbit} is the vector indicating the position of the prey, $r_1, r_2, r_3, r_4,$ and q are the random numbers, and $x(t)$ is the vector giving the current position of the hawk. LB and UB are the lower value and upper value of the positions. A hawk chosen randomly from the population is $x_{\text{rand}}(t)$, and the average position of the current hawk population is $x_m(t)$. Using Eq. 2, the average position is found [3].

$$x_m(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t) \quad (2)$$

The N value given in Eq. 2 indicates the number of hawks, while t is the number of iterations.

3.2.2 Transition from exploration to exploitation

After the hawks complete the exploration phase, they perform different attacks according to the energies of their prey. The decrease in the energy of the prey during escape is stated in Eq. 3 [48].

$$E = 2E_0\left(1 - \frac{t}{T}\right) \quad (3)$$

The E value in Eq. 3 indicates the energy of the escaped prey, the E_0 value the initial energy of the prey, and the T value the maximum number of repetitions.

3.2.3 Exploitation phase

This is the stage where the hawks make the surprise leap to their prey. Reacting to the surprise jump by attacking the hawk's prey, the rabbits try to escape. In response to these escapes of the rabbits, the Harris hawks employ different strategies. In the algorithm, these strategies are designed in 4 different ways.

The first strategy is called the "Soft besiege," where the Harris hawk tries to de-energize its prey by making deceptive leaps. ($r \geq 0.5, E \geq 0.5$). The soft besiege strategy is as in Eqs. 4 and 5 [48].

$$x(t+1) = \Delta x(t) - E|JX_{\text{rabbit}}(t) - x(t)| \quad (4)$$

$$\Delta x(t) = x_{\text{rabbit}}(t) - x(t) \quad (5)$$

Considering Eqs. 4 and 5, the escaped prey's chance of being caught is r, and the rabbit's energy E. The difference between the position in t iteration and the position of the rabbit is $\Delta x(t)$. The value indicated by J is a value that changes in each iteration to simulate the natural rabbit movement [48].

The second strategy has been called "Hard Besiege," where the hawk hardly flanks to throw its surprise claw against the prey's energy ($r \geq 0.5, |E| \leq 0.5$) that is greatly diminished. This situation is shown in Eq. 6 [48].

$$x(t+1) = x_{\text{rabbit}}(t) - E|\Delta x(t)| \quad (6)$$

The third strategy is "Soft besiege with progressive rapid dives," where the Prey has enough energy to escape. In other words, it makes a soft siege before the surprise jump compared to the previous siege. It is smarter than the hard siege step. In other words, it is shown in Eq. 7 that the hawks decide on their next step before making the soft siege [48].

$$Y = x_{\text{rabbit}}(t) - E|Jx_{\text{rabbit}}(t) - x(t)| \quad (7)$$

In order to decide whether this move made in the next step will be a good dive move, a comparison with the previous dive is made. If it is not suitable, a sudden dive is made. During this decision, a Levy Flight-based movement structure is used. It is given in Eq. 8.

$$Z = Y + S \times LF(D) \quad (8)$$

According to what is given in Eq. 8, the problem dimension is D. S is a random vector of size $1 \times D$. Y indicates the position of the prey relative to its decreasing energy. Z is the variable that decides whether the hawk will move to its prey. LF is the levy function and is found using Eq. 9 [48].

$$\text{LF}(x) = 0.01x\left(\frac{\mu x\sigma}{|\mu|^{\frac{1}{\beta}}}\right), \sigma = \left[\frac{\Gamma(1+\beta)x\sin(\frac{\pi\beta}{2})}{\Gamma\left(\frac{1+\beta}{2}\right)x\beta x2^{\left(\frac{\beta-1}{2}\right)}}\right] \quad (9)$$

According to what is given in Eq. 9, u, v is a random number between (0.1), and β is 1,5. Equation 10 is used to update the positions of the hawks [48].

$$x(t+1) = \{Y \text{ if } F(Y) < f(x(t)) \quad (10)$$

$$x(t+1) = \{Z \text{ if } F(Z) < F(x(t)) \quad (11)$$

The Y and Z equality given in Eqs. 10 and 11 is found using Eqs. 7 and 8 [48].

The last strategy has been called "Hard besiege with progressive rapid dives," at which stage the prey does not have enough energy to escape. The falcon makes a fierce siege before surprise leaps to capture prey. In Eqs. 12 and 13, the case of hard siege is given [48].

$$x'(t+1) = \{Y' \text{ if } F(Y') < f(x(t)) \quad (12)$$

$$x'(t+1) = \{Z' \text{ if } F(Z') < F(x(t)) \quad (13)$$

Values Y and Z can be determined from equations

$$Y' = x_{\text{rabbit}}(t) - E|Jx_{\text{rabbit}}(t) - x_m(t)| \quad (14)$$

$$Z' = Y' + S \times LF(D) \quad (15)$$

3.3 Multivariate adaptive regression spline (MARS)

The linear regression model is very important in solving many problems. However, real-life problems often show a nonlinear structure. Linear models cannot represent this structure well. Nonparametric regression is used to characterize these structures [56, 57]. If the number of independent variables to be used in the model to be created is large, nonparametric regression forms are not both useful and cannot be easily interpreted. However, MARS, developed by Friedman in 1991, is a form of nonparametric regression. While this method is useful for fitting nonlinear multivariate functions, it does not have the disadvantages of problems with a large number of independent variables.

MARS, which is nonparametric and does not assume a functional relationship between dependent and independent variables, has been used in many engineering problems [58, 59]. Instead of a mathematical relationship, it creates a dynamic relationship between cause and effect variables [60]. It builds a flexible regression model using basis functions corresponding to different ranges of independent variable [57, 61]. The model consists of data-driven basis functions and the coefficients associated with these bases. It divides the independent variable values into regions and associates each region with a regression equation.

General MARS model is defined as [61]

$$Y = \beta_0 + \sum_{k=1}^{K} a_k \beta_k(X_t) + \varepsilon_i \quad (16)$$

where k is the node number, K is number of the basic function, X is independent variable, a_k is the coefficients of the Kth basic function, β_0 is an constant term and $\beta_k(X_t)$ denotes kth independent variable for the kth basic function.

As $k = 1, 2, ..K$ defined, the basic function has the following form [61]

$$B_m(x) = \prod_{t=1}^{L_m}\left[S_{l,m}(x_{v(1,m)} - k_{l,m})\right] \quad (17)$$

According to Eq. 17, L_m is the degree of interaction, $S_{l,m} \in [\pm 1]$, $k_{l,m}$ is the node value, and $x_{v(1,m)}$ is the argument value.

In the first stage of the MARS method, which consists of two steps, forward and backward, since a more complex model than desired is obtained with the forward step algorithm, the basic functions in the model are added sequentially with the backward step algorithm, which is the second stage, to reach the optimum model. The process ends when the number of BF (basic functions) is maximized. However, the model produced at this stage includes the BFs that contribute most or least to the overall performance and is therefore more complex and contains inaccurate terms. The backward step is applied to avoid overfitting by reducing the complexity of the model without disturbing the fit of the model obtained in the forward step to the data. At each step, it removes the BF s that lead to the smallest increase in the residual sum of squares, and finally, a best-predicted model is created [62]. This process is called pruning and is determined by the hyperparameter "nprune." By allowing different shapes of BF s and their interactions, MARS has the capacity to reliably track very complex data structures that are often hidden in high dimensions [63, 64]. Pruning is most commonly done with the generalized cross-validation technique. These operations protect against overfitting by reducing the complexity of the model. The degree of pruning and which functions or interactions to remove affect the performance of the model.

Another hyperparameter that controls the degree of the polynomial basis functions in the method using the expansion of piecewise linear basis functions is the "degree" parameter. The degree parameter allows the model to learn different relationships. Increasing the degree allows the model to learn more complex models, and decreasing the degree allows it to learn simpler relations. For these reasons, the pruning parameter and the degree have a significant impact on the performance of the MARS model.

4 Application of HHA and MARS methods to the anemia prediction problem

In this section, it is stated how the HHA and MARS methods are adapted to the anemia disease prediction problem.

4.1 Adaptation of HHA algorithm to anemia disease problem

Since 6 different blood parameters are used with the Harris hawks method, the feature vector is given as follows.

$$B = [B_1, B_2, B_3, B_4, B_5, B_6] \tag{18}$$

In order to detect anemia with HHA, the results were tested by modeling in 3 different forms. As shown in Table 2, Model-1HHA refers to linear form, Model-2HHA refers to quadratic form and Model-3HHA refers to exponential form.

According to Eq. 18, column 1 of the feature vector (B_1) is the coefficient of RBC in blood, column 2 (B_2) is the coefficient of HGB in blood, column 3 (B_3) is HCT, column 4 (B_4) is MCV, column 5 (B_5) is MCH, and column 6 (B_6) is MCHC. These coefficients represent the effect sizes of the parameters (RBC, HGB, HCT, MCV, MCH, and MCHC) on classification. The effects of the parameters on the classification success are optimized with HHA to obtain the most appropriate weight values.

Multiple linear regression model adapted to blood data is expressed in terms of combination of the anemia variables;

$$\begin{aligned} y &= B_0 + B_1 HBG + B_2 \mathrm{RBC} + B_3 \mathrm{MCH} + B_4 \mathrm{WBC} \\ &\quad + B_5 \mathrm{MCV} + B_6 HCT \\ &= B_0 + \sum_{i=1}^{k} B_i x_i \end{aligned} \tag{19}$$

Multiple quadratic regression model can be established as the following:

$$\begin{aligned} y &= B_0 + B_1 \mathrm{HBG} + B_2 \mathrm{RBC} + \ldots + B_6 \mathrm{HCT} \\ &+ B_7 \mathrm{HBG}^2 + B_8 . \mathrm{HBG}. RBC + B_9 . \mathrm{HBG}. \mathrm{MCH} + \ldots \\ &+ B_{12} . \mathrm{HBG}. \mathrm{HCT} + B_{13} \mathrm{RBC}^2 + B_{14} \mathrm{RBC}. \mathrm{MCH} + \ldots \\ &+ B_{17} \mathrm{RBC}. \mathrm{HCT} + B_{18} \mathrm{MCH}^2 + \ldots \\ &+ \ldots + B_{27} \mathrm{HCT}^2 \end{aligned} \tag{20}$$

Multiple exponential regression model can be established as the following:

$$\begin{aligned} y &= B_0 + B_1 e^{B_7 \mathrm{HBG}} + B_2 e^{B_8 \mathrm{RBC}} \\ &+ B_3 e^{B_9 \mathrm{MCH}} + B_4 e^{B_{10} \mathrm{WBC}} + B_5 e^{B_{11} \mathrm{MCHC}} + B_6 e^{B_{12} HCT} \end{aligned} \tag{21}$$

The y values in Eqs. 19, 20, and 21 are the anemia value. B_i, $0 \le i \le 6$, $0 < \mathrm{i} < 27$, and $0 \le i \le 12$ are the parameters to be determined for Eqs. 19–21, respectively, using Harris hawks algorithm.

The cost function in multiple linear form is expressed as

Table 2 Types of models to be applied HHA

Model name to apply HHA	Model type
Model-1HHA	Multiple linear form
Model-2HHA	Multiple quadratic form
Model-3HHA	Multiple exponential form

$$J(Q) = \frac{1}{N}\sum_{i=1}^{N}(Y - X_i B^T)^2 \quad (22)$$

where (X) is the input set, X_i patient records in the i. patient registration Y are the label values, and the number of patients in the dataset is N.

By considering the problem as data mining parameter optimization, the model parameters (B) are tried to be estimated with the help of the HHA using the dataset. The classification of the data is managed whether the anemia exists or not within the scope of the study.

Application stages of the HHA algorithm for Model-1HHA:

Step 1:

The population size (N) and the maximum number of iterations (T) are defined. (In the study, $N = 50$ and $T = 250$ were taken.)

Step 2: A matrix B is produced as much as the population size.

$$(-5 < B_0 < 5 \text{ and } -3 < B_1, B_2 \ldots B_N < 3)$$

Step 3 The objective function is calculated by considering Eq. 23.

$$J = \frac{1}{2}\left(B_0 + \sum_{i=1}^{k} y_i - B_i x_i\right)^2 \quad (23)$$

Step 4 Each row of matrix B is set as the *x*rabbit rabbit position. *x*rabbit is calculated according to the objective function. According to the number of iterations, the *x*rabbit position that minimizes the objective function is calculated.

Step 5 Equation 24 formulates the energy of the hunt.

$$E = 2E_0\left(1 - \frac{t}{T}\right) \quad (24)$$

The first energy of the hunt is compared with the objective function value, and then, this value is reduced according to the number of iterations, thus minimizing the objective function. Then, in the following steps, the parameter values (prey position) change according to the energy of the prey.

Step 6: The position vector is updated when the energy of the prey is greater than or equal to 1.

Step 7: If the energy of the prey is less than 1, there are four different strategy chances (exploitation phase).

Step 7.1: if ($|E| \geq 0.5$ and $r \geq 0.5$), *xrabbit* is updated using the soft besiege.

Step 7.2 if ($|E| \geq 0.5$ and $r < 0.5$), *xrabbit* is updated using step 6 (hard siege).

Step 7.3 if ($|E| < 0.5$ and $r \geq 0.5$), *xrabbit* is updated using the Soft besiege with progressive rapid dives.

Step 7.4 if ($|E| < 0.5$ and $r < 0.5$), using the Hard besiege with progressive rapid dives, *xrabbit* is updated.

Step 8 Repeat step 5 until the number of iterations.

Step 9 The hunt is located.

The accuracy and average accuracy value obtained at each floor by applying tenfold cross-validation for Model-1HHA are given in Table 3. The high accuracy achieved on each fold shows that the model is not affected by the unbalanced dataset.

The accuracy and average accuracy value obtained at each floor by applying tenfold cross-validation for Model-2HHA are given in Table 4. The high accuracy achieved on each fold shows that the model is not affected by the unbalanced dataset.

The accuracy and average accuracy value obtained at each floor by applying tenfold cross-validation for Model-3HHA are given in Table 5. The high accuracy achieved on each fold shows that the model is not affected by the unbalanced dataset.

The coefficients produced at each layer by the HHA method applied to linear, quadratic, and exponential form models are shown in Table 6, Tables 7, and 8.

As shown in Tables 6, 7, and 8, the overall model coefficients were created by averaging the weight values produced by Model-1HHA, Model-2HHA, and Model-3HHA at each fold and the average coefficients are presented in Tables 9, 10, and 11, respectively.

In Tables 9, 10 and 11, the HHA algorithm is run for the mathematical model specified in Eqs. 19, 20 and 21 and the weight values that minimize the fitness function given in Eq. 22 are calculated.

4.2 Adaptation of the MARS method to the problem of anemia

The 16–17 parameters of the anemia models were determined as defined in Table 12, with Model-1MARS being

Table 3 Accuracy values at each fold by Model-1HHA

	Accuracy	Resample
1	0.9865	Fold01
2	0.9904	Fold02
3	0.9904	Fold03
4	0.9878	Fold04
5	0.9884	Fold05
6	0.9872	Fold06
7	0.9910	Fold07
8	0.9846	Fold08
9	0.9885	Fold09
10	0.9942	Fold10
Average accuracy	0.9889	

Table 4 Accuracy values at each fold by Model-2HHA

	Accuracy	Resample
1	0.9872	Fold01
2	0.9897	Fold02
3	0.9891	Fold03
4	0.9891	Fold04
5	0.9936	Fold05
6	0.9885	Fold06
7	0.9936	Fold07
8	0.9891	Fold08
9	0.9885	Fold09
10	0.9942	Fold10
Average accuracy	0.9903	

Table 5 Accuracy values at each fold by Model-3HHA

	Accuracy	Resample
1	0.9872	Fold01
2	0.9897	Fold02
3	0.9878	Fold03
4	0.9910	Fold04
5	0.9852	Fold05
6	0.9872	Fold06
7	0.9936	Fold07
8	0.9859	Fold08
9	0.9904	Fold09
10	0.9705	Fold10
Average accuracy	0.9868	

the first-order model, Model-2MARS second-order model, and Model-3MARS being the model with the best pruning and degree values, respectively.

To examine the performance of the MARS model with different combinations of hyperparameters, tests were performed on three different models. For the cross-validation process, tenfold cross-validation was tried. During the experiment, the hyperparameters degree and nprune were tested with different values. In Model-1MARS, degree = 1 and nprune = 25, in Model-2MARS, degree = 2 and nprune = 25, and finally in Model-3MARS, nprune was set from 5 to 50 (5 : 5 : 50) and degree was set between 1 and 4 (1 : 4) in order to find the most successful parameter combinations. With these hyperparameter combinations, it is aimed to capture the behavior of the model in a wide range.

The accuracy and average accuracy value obtained at each step fold by applying tenfold cross-validation for Model-1MARS are given in Table 13. The high accuracy achieved on each fold shows that the model is not affected by the unbalanced dataset.

Before the pruning process, 10 basis functions were generated and the functions with low success rate were removed from the model. Finally, in order to test the effect of RBC, HGB, HCT, MCV, MCH, and MCHC on anemia, 9 basis functions generated by the MARS method for Model-1MARS and their weights are presented in Table 14.

The basic functions defined for the model with BFj, where j is the number of basic functions, are presented in Table 14. In line with this information, the Model-1MARS model is as follows:

$$\begin{aligned}\text{Model-1MARS} = {} & 23.60 - 225.31x\text{BF1} + 237.19x\text{BF2} \\ & - 46.91x\text{BF3} - 27.17x\text{BF4} - 8.86x\text{BF5} \\ & + 12.05x\text{BF6} + 46.61x\text{BF7} + 6.99x\text{BF8}\end{aligned} \tag{25}$$

When Eq. 25 is analyzed, it is observed that there is no interaction of the basis functions by taking degree = 1, and simple linear functions are formed. According to Table 14 and Eq. 25, the RBC and WBC parameters were pruned and removed from the model by the MARS method because of their low contribution to the model performance.

The accuracy and average accuracy value obtained at each fold by applying tenfold cross-validation for Model-2

Table 6 Attribute weights at each fold by Model-1HHA

Attribute	Fold01	Fold02	Fold03	Fold04	Fold05	Fold06	Fold07	Fold08	Fold09	Fold10
B_0	1.722	3.750	2.949	1.538	2.307	1.578	1.041	0.655	2.868	3.650
B_1	− 1.999	− 1.752	− 1.923	− 1.083	− 0.329	− 0.890	− 0.135	− 0.595	− 1.897	0.255
B_2	− 1.104	− 2.000	− 1.648	− 1.055	− 1.422	− 1.991	− 1.152	0.628	− 2.000	− 2.000
B_3	0.382	− 1.324	− 1.686	− 0.479	− 1.999	− 1.008	− 0.013	− 1.642	− 1.983	− 1.399
B_4	− 0.081	− 0.856	0.981	0.470	1.450	2.000	0.335	1.998	1.863	− 1.839
B_5	− 0.664	− 0.738	− 1.535	− 0.857	− 1.167	− 2.000	− 0.334	− 1.502	− 1.942	1.454
B_6	− 0.050	− 1.058	− 0.484	0.266	− 0.849	0.734	0.152	− 0.484	− 0.482	− 1.794

Table 7 Attribute weights at each fold by Model-2HHA

Attribute	Fold01	Fold02	Fold03	Fold04	Fold05	Fold06	Fold07	Fold08	Fold09	Fold10
B_0	2.592	4.999	2.558	4.998	2.599	4.453	3.097	4.983	4.997	4.199
B_1	0.772	− 2.000	− 1.283	1.458	0.098	− 0286	2.000	0.571	− 2.000	− 1.939
B_2	− 1.280	1.999	0.475	− 1.991	0.475	1.617	− 1.259	1.788	1.994	− 0.466
B_3	− 2.000	− 1.739	− 1.925	− 1.999	0.666	− 0.646	0.502	− 0.757	− 2.000	0.903
B_4	1.695	− 1.470	1.661	0.409	1.582	0.169	− 0.305	− 1.381	− 1.607	1.872
B_5	0.742	− 0.608	1.174	1.432	− 0.444	− 2.000	1.502	− 0.864	0.213	2.000
B_6	− 0.147	0.748	− 1.279	0.236	− 0.682	− 2.000	1.303	1.675	1.265	− 2.000
B_7	1.218	4.337	2.149	− 3.651	− 0.517	− 2.514	− 2.434	− 4.812	4.688	− 3.779
B_8	2.063	− 3.428	− 3.542	− 3.883	0.419	4.536	− 0.914	− 4.995	0.348	2.959
B_9	− 0.794	− 3.842	4.949	4.028	0.399	− 0.643	− 0.849	0.209	− 5.000	3.011
B_{10}	− 1.835	− 1.254	4.999	− 4.995	− 1.645	− 1.394	− 1.284	4.413	− 4.053	− 4.698
B_{11}	− 0.882	4.999	− 4.989	4.745	0.758	1.757	0.213	3.230	1.047	0.836
B_{12}	− 4.956	1.676	− 5.000	− 1.886	1.110	2.981	− 1.284	− 3.839	4.725	1.309
B_{13}	− 2.407	− 1.927	− 5.000	− 0.055	− 0.824	− 2.967	− 0.595	− 2.590	− 4.606	− 2.024
B_{14}	0.655	− 5.000	− 3.300	0.666	− 0.946	− 0.222	− 2.441	− 0.628	3.044	− 3.401
B_{15}	− 1.561	− 1.830	0.302	− 0.349	0.042	− 5.000	− 1.065	− 0.666	− 5.000	− 2.321
B_{16}	2.653	0.142	1.303	− 3.253	1.757	3.828	− 3.763	− 0.653	− 5.000	− 1.902
B_{17}	0.363	0.152	− 4.449	3.248	− 4.359	− 1.604	− 2.899	− 2.734	4.216	− 1.608
B_{18}	− 4.589	2.821	0.644	− 4.205	− 4.104	− 4.172	− 2.455	− 2.396	− 0.729	− 2.258
B_{19}	− 0.778	1.663	4.671	0.448	− 0.141	− 2.282	1.014	− 2.842	− 1.294	− 3.673
B_{20}	− 3.498	− 1.765	− 2.326	1.271	− 0.253	− 1.777	− 3.009	4.271	− 0.052	− 1.215
B_{21}	2.022	− 3.605	− 0.100	− 4.993	− 4.849	1.752	0.542	0.284	− 5.000	2.399
B_{22}	4.948	− 0.255	− 4.543	0.404	0.759	2.715	1.509	− 3.341	4.997	2.817
B_{23}	− 0.904	− 1.492	3.269	0.986	− 2.551	3.587	1.274	2.025	2.795	− 5.000
B_{24}	3.928	− 0.295	3.869	− 2.918	− 0.161	− 5.000	1.798	− 1.156	− 5.000	1.608
B_{25}	− 5.000	3.874	− 4.446	− 3.688	0.656	0.273	− 1.551	1.181	− 0.170	1.165
B_{26}	− 3.861	− 0.743	1.148	1.331	− 0.627	− 2.029	− 0.241	0.853	− 1.232	2.159
B_{27}	− 2.366	− 3.293	2.216	− 1.917	2.949	1.309	− 1.947	− 4.890	− 1.885	− 3.740

MARS are given in Table 15. The high accuracy achieved on each fold shows that the model is not affected by the unbalanced dataset.

With Model-2MARS, 17 basis functions were generated before the pruning process and the functions with low success rates were removed from the model. Finally, in order to test the effect of RBC, HGB, HCT, MCV, MCH, and MCHC on anemia, 8 basis functions generated by the MARS method for Model-2MARS and their weights are presented in Table 16.

In line with the information in Table 16, the Model-2MARS model is as follows:

$$\begin{aligned}\text{Model-2MARS} = {} & 11.7 - 4.5x\text{BF1} - 585.3x\text{BF2} \\ & + 626.3x\text{BF3} - 45.5x\text{BF4} + 7.2x\text{BF5} \\ & + 294.1x\text{BF6} - 268.2x\text{BF7}\end{aligned} \tag{26}$$

When Eq. 26 is analyzed, the maximum interaction level of the basis functions is set to 2 by taking degree = 1. This shows that the model's basis functions are linear and quadratic functions as shown in Table 16.

BF1, BF2, BF3, and BF4 form linear functions; BF5, BF6, and BF7 are quadratic functions formed as the product of two basis functions. According to Table 16 and Eq. 26, MCH, WBC, and MCHC parameters were removed from the model by the MARS method because of their low contribution to the model's performance success.

The accuracy and average accuracy value obtained at each fold by applying tenfold cross-validation for Model-3MARS are given in Table 17. The high accuracy achieved on each fold shows that the model is not affected by the unbalanced dataset.

For Model-3MARS, degree = 1:4 and nprune = seq(5:5:50). As shown in Table 18, a grid containing the

Table 8 Attribute weights at each fold by model-3HHA

Attribute	Fold01	Fold02	Fold03	Fold04	Fold05	Fold06	Fold07	Fold08	Fold09	Fold10
B_0	− 1.414	− 2.861	− 0.602	− 3.846	− 1.844	− 0.237	− 5.000	− 2.468	− 2.936	0.422
B_1	1.188	2.369	− 0.199	3.000	0.239	0.845	3.000	2.992	2.038	1.032
B_2	1.052	− 2.210	3.000	2.568	2.489	2.123	2.849	2.976	2.638	2.199
B_3	2.551	3.000	1.697	2.299	3.000	− 0.289	2.458	− 2.694	3.000	1.374
B_4	0.011	− 0.454	− 0.915	− 1.916	3.000	0.416	0.096	1.215	− 0.821	− 0.718
B_5	2.828	2.681	1.954	2.077	− 0.570	0.930	2.234	0.526	3.000	1.740
B_6	2.686	2.651	2.592	− 0.304	1.515	− 0.039	1.707	0.105	3.000	− 1.708
B_7	− 1.031	− 6.698	− 0.209	− 1.137	− 5.180	− 2.471	− 2.249	− 0.686	− 4.165	− 2.984
B_8	− 6.219	− 7.853	− 3.123	− 3.479	− 1.978	− 5.013	− 2.044	− 6.420	− 4.991	− 9.964
B_9	− 2.412	− 2.309	− 0.997	0.000	− 1.184	− 5.659	− 3.276	− 8.317	− 2.613	− 9.430
B_{10}	− 8.639	− 3.295	− 1.610	− 0.590	− 4.364	− 4.829	2.812	− 2.770	0.000	− 0.130
B_{11}	− 9.964	− 0.932	− 9.982	− 2.354	− 0.438	− 4.623	− 1.623	− 3.421	− 0.637	− 4.692
B_{12}	− 6.784	− 6.477	− 9.982	− 2.246	− 5.772	− 2.996	− 0.869	− 2.493	− 6.246	− 9.945

Table 9 Average attribute weights by Model-1HHA

Attribute	Weight
B_0	2.206
B_1	− 1.036
B_2	− 1.374
B_3	− 1.115
B_4	0.632
B_5	− 0.862
B_6	− 0.405

Table 10 Average attribute weights by Model-2HHA

Attribute	Weight
B_0	3.947
B_1	− 0.261
B_2	− 0.335
B_3	− 0.90
B_4	0.263
B_5	0.315
B_6	− 0.088
B_7	− 0.531
B_8	− 0.644
B_9	0.147
B_{10}	− 1.174
B_{11}	1.172
B_{12}	− 0.587
B_{13}	− 2.3
B_{14}	− 1.157
B_{15}	− 1.745
B_{16}	− 0.489
B_{17}	− 0.969
B_{18}	− 2.144
B_{19}	− 0.321
B_{20}	− 0.836
B_{21}	− 1.155
B_{22}	1.001
B_{23}	0.399
B_{24}	− 0.333
B_{25}	− 0.771
B_{26}	− 0.324
B_{27}	− 1.356

performances of the hyperparameters according to the values in these ranges was created.

As a result of the tests, 17 basis functions were generated by Model-3MARS before the pruning process and the functions with low success rate with pruning process were removed from the model. Finally, in order to test the effect of RBC, HGB, HCT, MCV, MCH, and MCHC on anemia, 5 basis functions generated by the MARS method for Model-3MARS and their weights are presented in Table 19.

In line with the information in Table 19, the Model-3MARS model is as follows:

$$\text{Model} - 3\text{MARS} = 43.5 - 1107.3x\text{BF1} + 671.4x\text{BF2} + 419.9x\text{BF3} - 123.8x\text{BF4} \quad (27)$$

When Eq. 27 is analyzed, the maximum interaction level of the basis functions is set to 2 by taking degree = 2 and nprune = 5. This shows that the model's basis functions are linear and quadratic functions as shown in Table 19. BF1, BF2, and BF3 are linear functions, while

Table 11 Average attribute weights by Model-2HHA

Attribute	Weight
B_0	− 2.077
B_1	1.651
B_2	1.528
B_3	1.640
B_4	− 0.009
B_5	1.740
B_6	1.221
B_7	− 2.681
B_8	− 5.108
B_9	− 3.620
B_{10}	− 2.904
B_{11}	− 3.867
B_{12}	− 5.381

Table 12 Model types to which the MARS method is applied

Model Name applied to MARS	Model type
Model-1MARS	Degree = 1(linear) AND nprune = 25
Model-2MARS	Degree = 2 (quadratic) nprune = 25
Model-3MARS	Degree = 1:4 AND nprune(5.5:50)(quadratic)

Table 13 Accuracy values at each fold by Model-1MARS

	Accuracy	Resample
1	1.000	Fold01
2	0.9885	Fold02
3	0.9942	Fold03
4	0.9827	Fold04
5	0.9827	Fold05
6	0.9828	Fold06
7	0.9942	Fold07
8	0.9942	Fold08
9	0.9884	Fold09
10	0.9942	Fold10
Average accuracy	0.9873	

BF4 is a quadratic function which is the product of two basis functions. According to Table 19 and Eq. 27, MCH, WBC, HCT, and MCHC parameters were pruned and removed from the model by the MARS method in the pruning phase since their contribution to the model performance was low.

Table 14 Degree = 1 Estimated results of the MARS model

	Basic function	Weight
	Constant	23.60
BF1	max(0, HBG—0.480916)	− 225.31
BF2	max(0, HBG—0.618321)	237.19
BF3	max(0, 0.679389-HBG)	− 46.91
BF4	max(0, HBG-0.679389)	− 27.17
BF5	max(0, HCT-0.491623)	− 8.86
BF6	max(0, HCT-0.552046)	12.05
BF7	max(0, 0.284615—MCHC)	46.61
BF8	max(0, MCHC-0.284615)	6.99

Table 15 Accuracy values at each fold by Model-2MARS

	Accuracy	Resample
1	0.9885	Fold01
2	0.9653	Fold02
3	1.0000	Fold03
4	0.9942	Fold04
5	0.9885	Fold05
6	0.9884	Fold06
7	0.9942	Fold07
8	0.9942	Fold08
9	0.9942	Fold09
10	0.9942	Fold10
Average accuracy	0.9902	

When a new patient data is to be analyzed using Eqs. 25–27, classification can be made by entering the necessary information in the BF*j* basic functions.

5 Evaluation

In the study, two classes were expressed as anemia (1) and non-anemia/healthy (0) individuals. A tenfold cross-validation method was used for all proposed models. With this method, the dataset is divided into 10 different subsets, and each time one group is used as a test set, while the other group is used for training. This approach enables an evaluation process where each subset serves as a test set once and all combinations are tested and the results averaged. This way, the class imbalance in the dataset does not cause

Table 16 Degree = 2 Estimated results of the MARS model

	Basic function	Weight
	Constant	11.7
BF1	max(0, HGB—0.51145)	− 4.5
BF2	max(0, HGB—0.541985)	− 585.3
BF3	max(0, HGB—572,519)	626.3
BF4	max(0, HGB—59,542)	− 45.5
BF5	max(0, RBC-0.212613) * max(0, 0.679389—HGB)	7.2
BF6	max(0, HGB-0.51145) * max(0, 0.612469—HCT)	294.1
BF7	max(0, 0.679389-HGB) * max(0, HCT-0.502609)	− 268.2

Table 17 Accuracy values at each fold by Model-3MARS

	Accuracy	Resample
1	0.9884	Fold01
2	0.9885	Fold02
3	0.9885	Fold03
4	0.9884	Fold04
5	1.0000	Fold05
6	0.9942	Fold06
7	0.9942	Fold07
8	0.9942	Fold08
9	0.9942	Fold09
10	1.0000	Fold10
Average accuracy	0.9906	

any problems during model learning and performance evaluation. In other words, since there are more non-anemia records in the dataset, the model may learn this class better, while it may tend to misclassify the anemia class with fewer records in the anemia class. However, with this method, where the representation of each class in the training and test sets is compatible with the proportion in the overall dataset, the effect of such situations is minimized.

The performances of the classification methods were calculated according to ROC analysis metrics. These metrics are methods that reveal how well the model performs in predictions in order to learn the performance of the results obtained by various methods. It is frequently used in data mining applications [65].

Confusion matrix for ROC analysis is shown in Table 20.

The basic equations for ROC analysis are shown in Eqs. 28–33.

Table 18 Accuracy values at each fold by the grid containing hyperparameters Degree = 1:4 and nprune = seq(5:5:50)

	Degree	nprune	Accuracy
1	1	5	0.9862
2	2	5	0.9931
3	3	5	0.9913
4	4	5	0.9913
5	1	10	0.9873
6	2	10	0.9896
7	3	10	0.9925
8	4	10	0.9925
9	1	15	0.9873
10	2	15	0.9902
11	3	15	0.9925
12	4	15	0.9925
13	1	20	0.9873
14	2	20	0.9902
15	3	20	0.9925
16	4	20	0.9925
17	1	25	0.9873
18	2	25	0.9902
19	3	25	0.9925
20	4	25	0.9925
21	1	30	0.9873
22	2	30	0.9902
23	3	30	0.9925
24	4	30	0.9925
25	1	35	0.9873
26	2	35	0.9902
27	3	35	0.9925
28	4	35	0.9925
29	1	40	0.9873
30	2	40	0.9902
31	3	40	0.9925
32	4	40	0.9925
33	1	45	0.9873
34	2	45	0.9902
35	3	45	0.9925
36	4	45	0.9925
37	1	50	0.9873
38	2	50	0.9902
39	3	50	0.9925
40	4	50	0.9925

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{28}$$

Table 19 Degree = 1:4 and nprune = seq(5:5:50) estimated results of the MARS model

	Basic function	Weight
	Constant	43.5
BF1	max(0, HGB—0.51145)	− 1107.3
BF2	max(0, HGB—0.541985)	671.4
BF3	max(0, HGB—572,519)	419.9
BF4	max(0, RBC-0.212613) * max(0, 0.679389—HGB)	− 123.8

Table 20 Confusion matrix for ROC

		Predicted	
		0	1
Actual	0	TN	FP
	1	FN	TP

$$\text{Recall} - \text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{29}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{30}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{31}$$

$$\text{F1} - \text{Score} = \frac{2\text{Precision.Recall}}{\text{Precision} + \text{Recall}} \tag{32}$$

$$\text{AUC} = \frac{\text{TPR} - \text{TNR}}{2} \tag{33}$$

The purpose of ROC analysis is to compare the performance of the results obtained by various methods and to evaluate the results in terms of sensitivity, specificity, precision, F1-score, AUC, and accuracy [66]. The ROC parameters used in the analysis are TP, TN, FP, and FN. TP (true positive) and TN (true negative) indicate the correct prediction of anemia as a result of the classification used. FP (false positive) and FN (false negative) represent the number of false predictions.

Selecting the appropriate metric for the model is very important for obtaining the desired results. Accuracy, AUC, and F1-score are parameters often used to evaluate model performance. All three are used to evaluate how well a model performs. Accuracy is the most popular metric that determines the percentage of correct predictions. AUC compares the relationship between the true-positive rate (TPR) and false-positive rate (FPR) at different thresholds. Data scientists try to achieve the highest TPR while maintaining the lowest FPR, indicating their success in making correct predictions. In unevenly distributed datasets, it is not enough to measure model success with the accuracy metric alone. F1-score is one of the most widely used metrics for unbalanced datasets [67, 68]. F-score is a measure based on precision and recall values. That is, if recall is high, precision is usually low, and vice versa. At the same time, a high recall value means that a large proportion of instances in the minority class are correctly predicted, while a high precision value indicates that there is a high probability that the predicted minority instances actually belong to the minority class. Higher values for both precision and recall lead to a higher F1-score, indicating that there is not a large difference between the precision and recall values [68, 69].

Although AUC is a very common measure for imbalanced data problems, the F-measure is more suitable for such cases. This is because the minority class is more critical than the majority class, and the F-measure is an indication that the desired method classifies samples in the minority class with a higher accuracy and a lower misclassification rate. In other words, the AUC measure evaluates the overall accuracy of the classifier in both the majority and minority classes, while the F-measure focuses only on the accuracy of the classifier in the minority class [68–70].

Since the dataset used in the study is an imbalanced dataset, AUC and F1-score performance comparison is made in the next section, but since other studies in the literature produce results based on the accuracy metric, other ROC metrics are also presented.

6 Experimental results of HHA and MARS

In this study, anemia disease classification was performed with 2 different optimization algorithms for patient data with blood values and results. The data used in the study were classified using the two methods given in Tables 2 and 12, and the performance of each method was tested with three different models. Classification was performed on linear, quadratic, and exponential models for the HHA method, and on models with different pruning coefficients such as first-order pruning coefficient 25, second-order pruning coefficient 25, and degree values between 1 and 4 for the MARS method.

The high accuracy values in Tables 3, 4 and 5 for the HHA method and Tables 13, 15 and 17 for the MARS method show that the accuracy of the proposed methods is high in each cluster in the dataset divided into 10 different

subsets. This shows that the proposed methods are not affected by dataset imbalance.

This section presents the confusion matrices and ROC performance analysis of the methods obtained for the six models. The confusion matrices of both Model-1 and Model-3HHA algorithms and Model-1 and Model-3MARS methods are documented in Figs. 1, 2, 3, 4, 5 and 6, respectively, and ROC performance values are calculated from them.

When the confusion matrices in Figs. 1, 2, 3, 4, 5 and 6 are evaluated, it is seen that a total of 13 patients with Model-1HHA, 12 patients with Model-2HHA, 316 patients with Model-3HHA, 14 patients with Model-1MARS, 12 patients with Model-2MARS and 11 patients with Model-3MARS could not be classified correctly.

Table 21 shows the ROC performance analyses made as a result of testing six different models.

In the study, in order to better model the relationship between anemia disease parameters, 6 different tests were performed at different degrees and emphasizing the interaction of parameters with each other. In the research conducted on different models, it was analyzed that the classical method MARS and the metaheuristic method HHA showed extremely significant success in anemia disease classification when the accuracy metric was considered. However, since it is necessary to evaluate model performance metrics such as precision, recall, F1-score, and AUC in addition to the accuracy metric to determine model success in non-uniformly distributed datasets, these metrics were also analyzed.

For Model-1 and Model-3 HHA, the precision metric, which is the ability of the classifier not to label a healthy record as anemia, is high for all 3 models. Recall, F1-score, and AUC metrics are high in Model-1 and Model-2, but

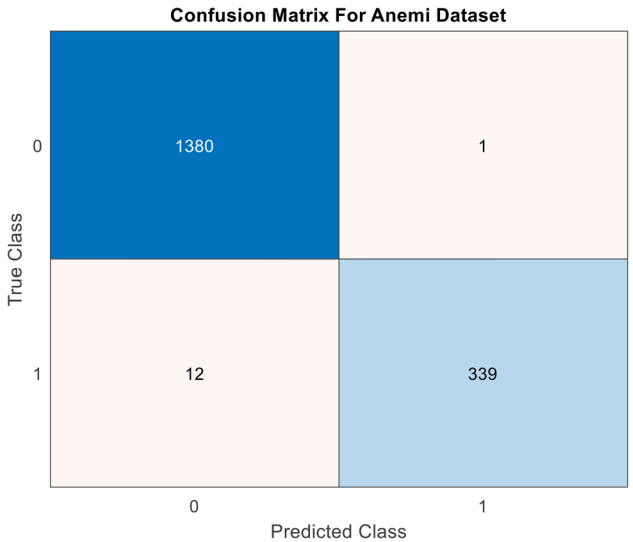

Fig. 1 Confusion matrix of the Model-1HHA

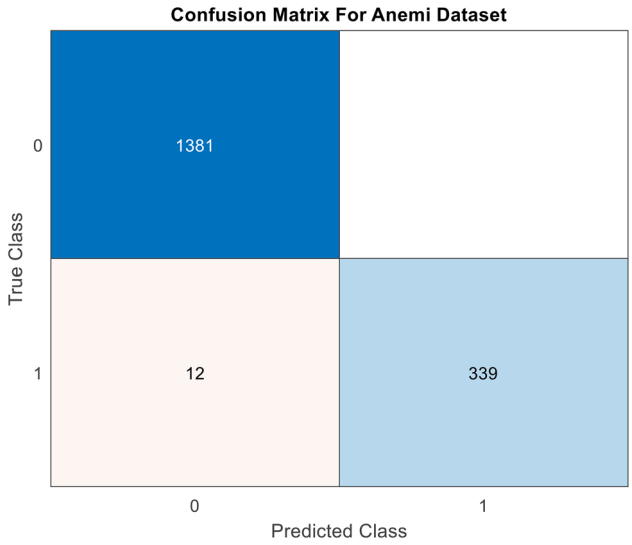

Fig. 2 Confusion matrix of the Model-2HHA

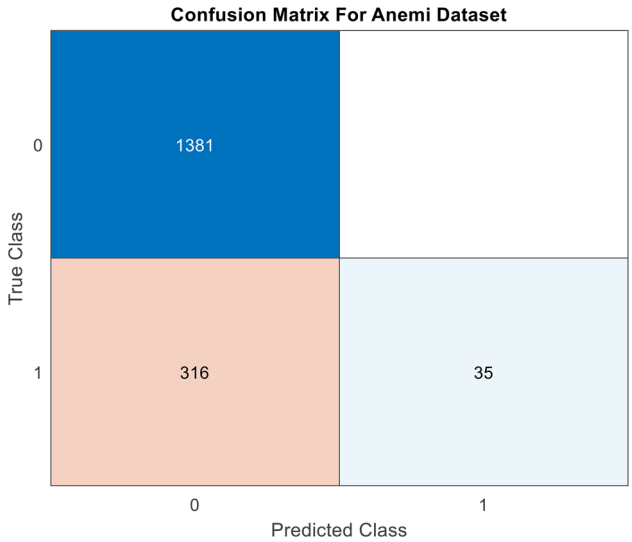

Fig. 3 Confusion matrix of the Model-3HHA

low in Model-3. In the dataset where the number of anemia patient records is low, class-based achievements are important instead of overall accuracy achievement. In other words, while the accuracy value is 81.76%, the recall metric, which is the ratio of all correctly predicted anemia patient records to anemia patient records, is low in Model-3 because of the low number of anemia patient records. The F1-score, which is also the weighted average of precision and recall, takes into account the misclassified anemia records from the precision score and the misclassified healthy data records from the recall score. The AUC metric, which evaluates the accuracy of the classifier in both the minority and majority classes, and the F1-score metric, which evaluates the accuracy of the classifier only in the minority class, also produced high results for Model-

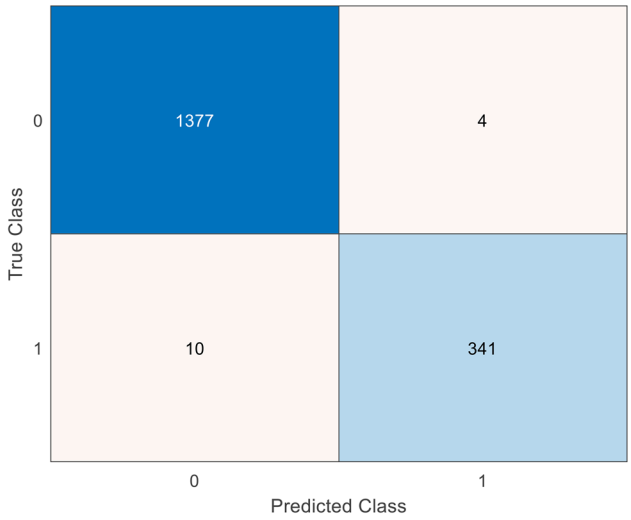

Fig. 4 Confusion matrix of the Model-1MARS

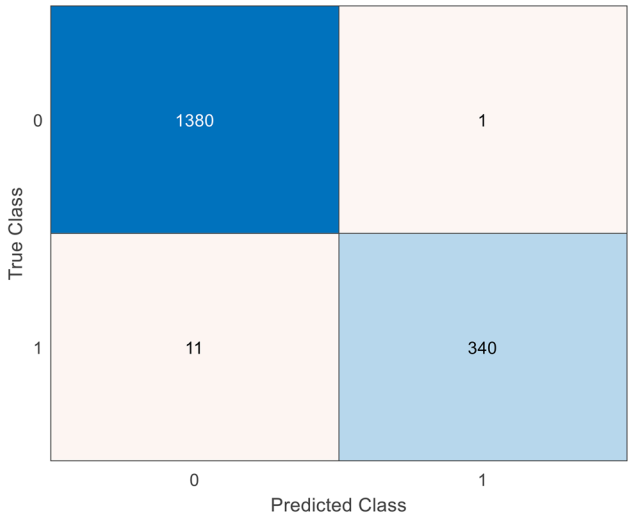

Fig. 5 Confusion matrix of the Model-2MARS

1 and Model-2, but poor results for Model-3, which shows exponential behavior.

When the results of Model-1 and Model-3MARS where the MARS method is applied are evaluated, it is seen that first-order Model-1 and second-order Model-2 show successful results when nprune = 25 is selected. In the tests performed for Model-3, which finds the best degree and prune hyperparameters to classify the anemia dataset with the MARS method, degree = 2 nprune = 5 was found. Again, it is seen that the success of Model-3 is high.

The cost function changes of linear, exponential, and quadratic HHA models using the mean squared error (MSE) function are shown in Figs. 7, 8, and 9. The reason why there is no continuous decrease in the graph drawn for Model-1 in Fig. 7 is that the cost curves are calculated by averaging the coefficients. Figure 7 shows that the linear

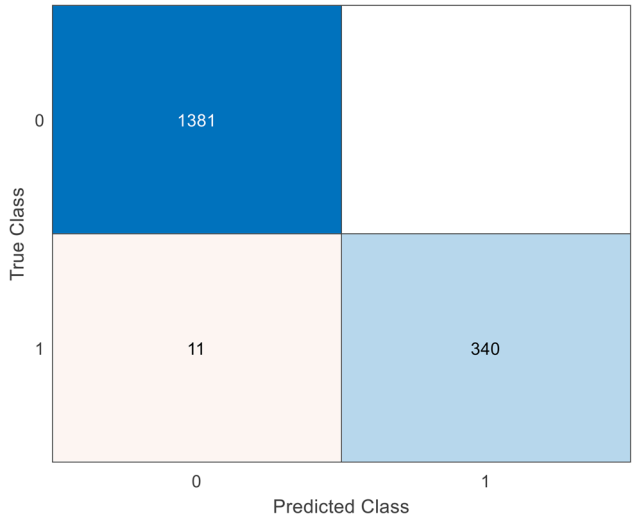

Fig. 6 Confusion matrix of the Model-3MARS

HHA model converges to the optimum result at the end of the 250th iteration.

The reason why there is no continuous decrease in Fig. 8 according to the cost values plotted by averaging the coefficients of the parameters at each iteration is that the cost values are calculated again by averaging the coefficients as in Model-1HHA. Looking at Fig. 8, the HHA model, which shows linear behavior according to the coefficients to be optimized, converges to the optimum result.

In Fig. 9, for the exponential form, the cost values are plotted by averaging the coefficients of the parameters at each iteration, resulting in a fluctuating graph. The reason for the lack of a continuous decrease on the graph is that the coefficients are obtained by averaging in each iteration as in Model-1 and Model-2. It is concluded that the reduction from 130 to 13 parameters for the exponential form is not suitable for a model with exponential behavior. In other words, averaging the parameters and obtaining a general parameter do not give healthy results in some cases.

These results show that the relationship between the parameters and the contribution of the model to the anemia disease classification problem is significant.

7 Conclusion

Harris hawks algorithm and MARS algorithm were used to classify the data in the database where each patient data contain 6 different blood components including RBC, HGB, HCT, MCV, MCH and MCHC blood values and the corresponding anemia outcome information. During classification, the dataset was divided into 10 different subsets

Table 21 ROC performance analyses

Model name	Accuracy	Precision	Recall	F1-score	AUC
Model-1 HHA	%99.25	%99.71	%96.58	%98.12	98.25
Model-2 HHA	%99.31	%100	%96.58	%98.26	98.29
Model-3HHA	%81.76	%100	%9.972	%18.13	54.99
Model-1 MARS	%99.19	%98.84	%97.15	%97.99	98.43
Model-2 MARS	%99.31	%98.84	%97.15	%97.99	98.40
Model-3 MARS	%99.36	%100	%96.87	%98.41	98.43

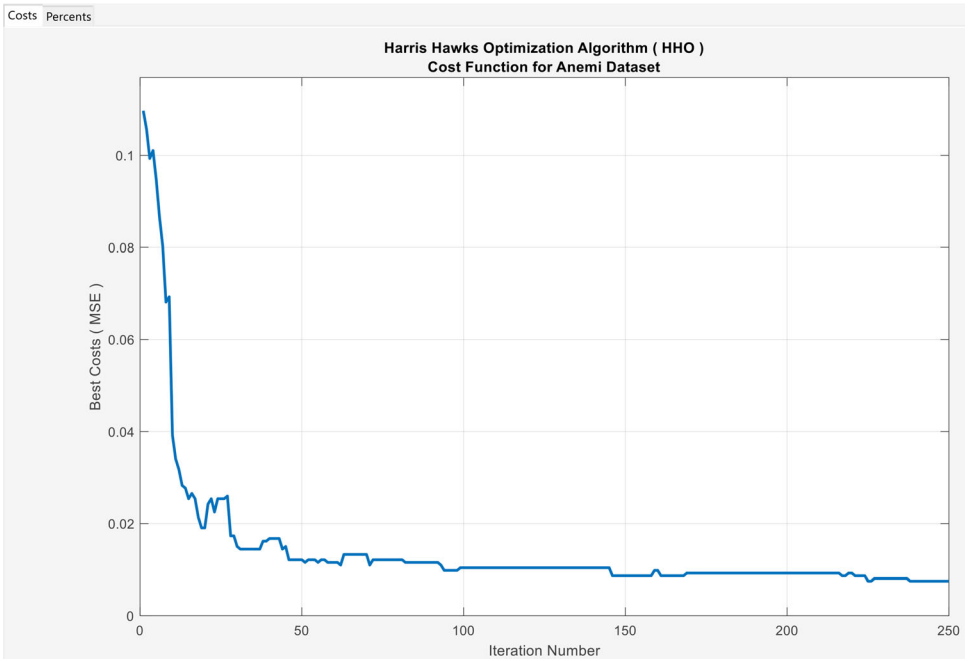

Fig. 7 The cost function of Model-1HHA

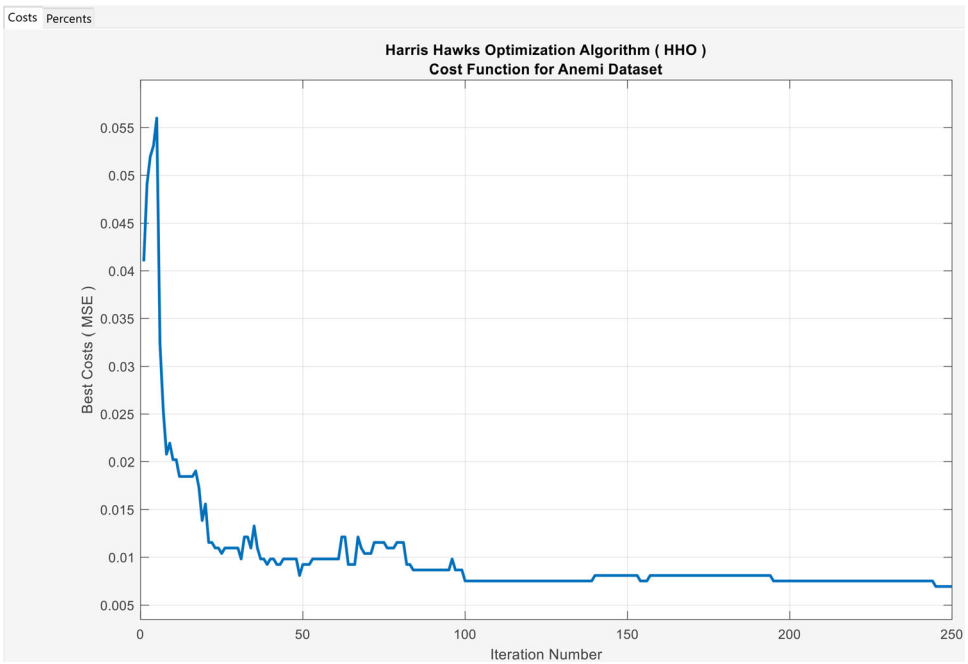

Fig. 8 The cost function of Model-2HHA

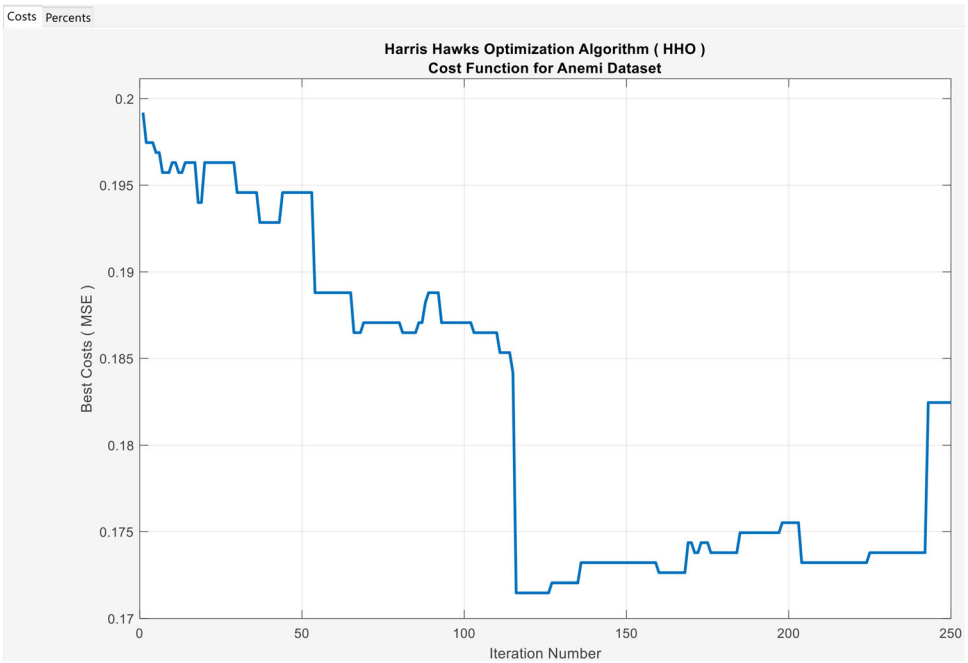

Fig. 9 The cost function of Model-3HHA

Table 22 Related studies using different datasets similar to our study

Data	Methods	References	Accuracy
RBC, HGB HCT, MCV MCH, MCHC, RDW	LR, k-NN, SVM, Extreme Learning Machine and Regularized Extreme Learning Machine	[31]	%95.59
MCV MCH, MCHC, HGB, RBC	ANN and an adaptive neuro-fuzzy inference system (ANFIS)	[36]	%96.29
MCV, RBC, HGB, HCT, MCH, MCHT	Clonal selection algorithm, Gini algorithm, FFN, PNN	[71]	%98.73
RBC, MCV, HCT, HGB	k-NN, CDTA, ANN	[43]	%78.31
Age, Gender, MCV, HCT, HGB, MCHC, RDW	Naïve Bayes, neural network J48 decision tree algorithms, and SVM	[40]	%93.75
HGB, HCT, MCH, MCV	RF, NB and CDTA	[72]	%96.09
Results for our proposed method			
Data	Methods	Accuracy	
HGB, RBC, MCH, WBC, MCHC, HCT	HHA Algorithm in Multiple Linear Form	%99.25	
HGB, RBC, MCH, WBC, MCHC, HCT	HHA Algorithm in Multiple Quadratic Form	%99.31	
HGB, RBC, MCH, WBC, MCHC, HCT	HHA Algorithm in Multiple Exponential Form	%81.76	
HGB, RBC, MCH, WBC, MCHC, HCT	First-order and nprune = 25 MARS Algorithm	%99.19	
HGB, RBC, MCH, WBC, MCHC, HCT	2nd Degree and nprune = 25MARS Algorithm	%99.31	
HGB, RBC, MCH, WBC, MCHC, HCT	2nd Degree and nprune = 5 MARS Algorithm	%99.36	

using the tenfold cross-validation method. In this process, each of the 10 subsets was tested once as a test set. The algorithms were analyzed on 6 different models: multilinear form HHA, multilinear quadratic form HHA, multi-exponential form HHA, MARS model with first-order pruning coefficient 25, MARS model with second-order pruning coefficient 25, and MARS model using the best degree and pruning coefficient.

As a result of the tests performed on the models applying the MARS method based on classical mathematical modeling, it was concluded that they performed well in anemia classification with an accuracy of 99.19%, 99.31%, and 99.36%, respectively, as shown in Table 22. In addition to the accuracy metric, it was also observed to perform well against the F1-score and AUC metrics, which

are commonly used to analyze imbalanced datasets (Table 21).

The results of the success of the parameter estimation method based on mathematical modeling using the HHA method on the anemia classification problem were 99.25%, 99.31%, and 81.76%, respectively. Thus, according to the coefficients to be optimized according to the average parameter approach, which obtains a general parameter that best summarizes the problem by averaging the parameter values produced at each floor in the cross-validation results, it is seen that the linear and quadratic model with linear behavior shows successful results, while the model in exponential form, which does not show linear behavior, shows lower success compared to other models.

When the results are evaluated, in the prediction process using both HHA and MARS, the targeted outputs were successfully achieved using 6 different blood components in a total of 1732 cases, 351 with anemia, and 1381 without anemia.

According to the performance results, the proposed algorithms classify anemia better than previous methods (Table 22). This shows that our proposed methods for the anemia classification problem have high performance in terms of accuracy, F1-score, and AUC performance. The high performance obtained and the high accuracy at each level of the dataset analyzed by tenfold cross-validation show that the proposed methods are not affected by the imbalance of the dataset.

The results of the study are expected to help medical students and doctors in the anemia classification problem. We believe that the classifier performances of our proposed models will contribute positively to the literature.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK). The authors did not receive support from any organization for the submitted work.

Data availability This research uses only publicly available anemia disease data from the Kaggle platform.

Declarations

Conflict of interest The authors have no relevant financial or nonfinancial interests to disclose.

References

1. De Benoist B, Cogswell M, Egli I, McLean E (2008) Worldwide prevalence of anaemia 1993–2005; WHO Global Database of anaemia
2. WHO. "Anaemia," World Health Organization. https://www.who.int/healthtopics/anaemia#tab=tab_1. Accessed 04 Oct 2023
3. Moraru L, Moldovanu S, Biswas A (2014) Optimization of breast lesion segmentation in texture feature space approach. Med Eng Phys 36(1):129–135
4. Dey N et al (2019) Social-group-optimization based tumor evaluation tool for clinical brain MRI of Flair/diffusion-weighted modality. Biocybern Biomed Eng 39(3):843–856
5. Sisodia D, Sisodia DS (2018) Prediction of diabetes using classification algorithms. Procedia Comput Sci 132:1578–1585
6. Thirunavukkarasu K, Singh AS, Rai P, Gupta S (2018) Classification of IRIS dataset using classification based KNN algorithm in supervised learning. İn: 2018 4th ınternational conference on computing communication and automation (ICCCA), IEEE, 2018, pp 1–4
7. Kuru K, Niranjan M, Tunca Y, Osvank E, Azim T (2014) Biomedical visual data analysis to build an intelligent diagnostic decision support system in medical genetics. Artif Intell Med 62(2):105–118
8. Bourouis A, Feham M, Hossain MA, Zhang L (2014) An intelligent mobile based decision support system for retinal disease diagnosis. Decis Supp Syst 59:341–350
9. Saba T et al (2019) Cloud-based decision support system for the detection and classification of malignant cells in breast cancer using breast cytology images. Microsc Res Tech 82(6):775–785
10. Borra S, Dey N, Bhattacharyya S, Bouhlel MS (2019) Intelligent decision support systems: applications in signal processing, vol 4. Walter de Gruyter GmbH & Co KG, Berlin
11. Kavas PÖ, Bozkurt MR, Kocayiğit İ, Bilgin C (2023) Machine learning-based medical decision support system for diagnosing HFpEF and HFrEF using PPG. Biomed Signal Process Control 79:104164
12. Uddin S, Khan A, Hossain ME, Moni MA (2019) Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 19(1):1–16
13. Kilicarslan S, Celik M, Sahin Ş (2021) Hybrid models based on genetic algorithm and deep learning algorithms for nutritional Anemia disease classification. Biomed Signal Process Control 63:102231
14. Yagmur N, Alagoz BB Modeling of first order plus time delay system dynamics with adaptive IIR filters based on gradient descent methods and performance analyses for different time delay cases. Pamukkale Univ J Eng Sci, vol 1000, no 1000
15. Duchi J, Hazan E, Singer Y (2011) Adaptive subgradient methods for online learning and stochastic optimization. J Mach Learn Res 12(7):2121–2159
16. Manogaran G, Lopez D (2018) Health data analytics using scalable logistic regression with stochastic gradient descent. Int J Adv Intell Paradig 10(1–2):118–132
17. Dixit M, Upadhyay N, Silakari S (2015) An exhaustive survey on nature inspired optimization algorithms. Int J Softw Eng Its Appl 9(4):91–104
18. Kumar SR, Singh KD (2021) Nature-inspired optimization algorithms: research direction and survey. arXiv preprint arXiv:2102.04013

19. Gundluru N et al (2022) Enhancement of detection of diabetic retinopathy using Harris hawks optimization with deep learning model. Comput Intell Neurosci. https://doi.org/10.1155/2022/8512469
20. Kumar A, Kabra G, Mussada EK, Dash MK, Rana PS (2019) Combined artificial bee colony algorithm and machine learning techniques for prediction of online consumer repurchase intention. Neural Comput Appl 31:877–890
21. Chou S-M, Lee T-S, Shao YE, Chen I-F (2004) Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. Expert Syst Appl 27(1):133–142
22. Bui DT et al (2019) A new intelligence approach based on GIS-based multivariate adaptive regression splines and metaheuristic optimization for predicting flash flood susceptible areas at high-frequency tropical typhoon area. J Hydrol 575:314–326
23. Goh ATC, Zhang Y, Zhang R, Zhang W, Xiao Y (2017) Evaluating stability of underground entry-type excavations using multivariate adaptive regression splines and logistic regression. Tunn Undergr Sp Technol 70:148–154
24. Deo RC, Kisi O, Singh VP (2017) Drought forecasting in eastern Australia using multivariate adaptive regression spline, least square support vector machine and M5Tree model. Atmos Res 184:149–175
25. Gill MK, Kaheil YH, Khalil A, McKee M, Bastidas L (2006) Multiobjective particle swarm optimization for parameter estimation in hydrology. Water Resour Res. https://doi.org/10.1029/2005WR004528
26. Küçükkülahli E, Erdoğmuş P, Polat K (2017) A hybrid approach to image segmentation: combination of BBO (Biogeography based optimization) and histogram based cluster estimation. İn: 2017 25th signal processing and communications applications conference (SIU), IEEE, 2017, pp 1–4
27. Yağmur N (2023) Üç Boyutlu Engelli Kübik Bir Ortamda Genetik Algoritma İle Robot Yol Planlamasında En Kısa Yol Bulma
28. Yağmur N, Alagöz BB (2019) Comparision of solutions of numerical gradient descent method and continous time gradient descent dynamics and lyapunov stability. İn: 2019 27th signal processing and communications applications conference (SIU), IEEE, pp 1–4
29. Özdemir D, Dörterler S (2022) An adaptive search equation-based artificial bee colony algorithm for transportation energy demand forecasting. Turkish J Electr Eng Comput Sci 30(4):1251–1268
30. Ahmad A, Alzaidi K, Sari M, Uslu H (2023) Prediction of anemia with a particle swarm optimization-based approach. Int J. Optim. Control Theor. Appl. 13(2):214–223
31. Çil B, Ayyıldız H, Tuncer T (2020) Discrimination of β-thalassemia and iron deficiency anemia through extreme learning machine and regularized extreme learning machine based decision support system. Med Hypotheses 138:109611
32. Wongseree W, Chaiyaratana N, Vichittumaros K, Winichagoon P, Fucharoen S (2007) Thalassaemia classification by neural networks and genetic programming. Inf Sci (NY) 177(3):771–786
33. Dogan S, Turkoglu I (2008) Iron-deficiency anemia detection from hematology parameters by using decision trees. Int J Sci Technol 3(1):85–92
34. Sanap SA, Nagori M, Kshirsagar V (2011) Classification of anemia using data mining techniques. İn: International conference on swarm, evolutionary, and memetic computing, Springer, Berlin, pp 113–121
35. Allahverdi N, Tunali A, Işik H, Kahramanli H (2011) A Takagi-Sugeno type neuro-fuzzy network for determining child anemia. Expert Syst Appl 38(6):7415–7418
36. Azarkhish I, Raoufy MR, Gharibzadeh S (2012) Artificial intelligence models for predicting iron deficiency anemia and iron serum level based on accessible laboratory data. J Med Syst 36:2057–2061
37. Yılmaz Z, Bozkurt MR (2012) Determination of women iron deficiency anemia using neural networks. J Med Syst 36:2941–2945
38. Setsirichok D et al (2012) Classification of complete blood count and haemoglobin typing data by a C4. 5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening. Biomed Signal Process Control 7(2):202–212
39. Akrimi JA, Rahimahmad A, George LE (2013) Review of machine learning techniques in Anemia recognition. Int J Sci Res (IJSR), India Online ISSN, pp 2319–7064
40. Abdullah M, Al-Asmari S (2017) Anemia types prediction based on data mining classification algorithms. In: Communication, management and ınformation technology, Alencar pp 615–621
41. Shahin AI, Guo Y, Amin KM, Sharawi AA (2019) White blood cells identification system based on convolutional deep neural learning networks. Comput Methods Programs Biomed 168:69–80
42. Dimauro G, Caivano D, Girardi F (2018) A new method and a non-invasive device to estimate anemia based on digital images of the conjunctiva. IEEE Access 6:46968–46975
43. İlaslaner T, Güven A (2019) Investigation of the effects biochemistry on ıron deficiency anemia. İn: 2019 medical technologies congress (TIPTEKNO), IEEE, 2019, pp 1–4
44. Khan JR, Chowdhury S, Islam H, Raheem E (2019) Machine learning algorithms to predict the childhood anemia in Bangladesh. J Data Sci 17(1):195–218
45. El-Kenawy EMT (2019) A machine learning model for hemoglobin estimation and anemia classification. Int J Comput Sci Inf Secur 17(2):100–108
46. Yıldız TK, Yurtay N, Öneç B (2021) Classifying anemia types using artificial learning methods. Eng Sci Technol Int J 24(1):50–70
47. Vohra R, Dudyala AK, Pahareeya J, Hussain A (2022) Decision rules generation using decision tree classifier and their optimization for anemia classification. In: ınventive computation and ınformation technologies: proceedings of ICICIT 2021, Springer, Berlin, pp 721–737
48. Heidari AA, Mirjalili S, Faris H, Aljarah I, Mafarja M, Chen H (2019) Harris hawks optimization: algorithm and applications. Futur Gener Comput Syst 97:849–872
49. Naeijian M, Rahimnejad A, Ebrahimi SM, Pourmousa N, Gadsden SA (2021) Parameter estimation of PV solar cells and modules using Whippy Harris hawks optimization algorithm. Energy Rep 7:4047–4063
50. Akdağ O, Abdullah A, Yeroglu C (2020) Harris Şahini optimizasyon Algoritması ile Aktif Güç Kayıplarının minimizasyonu. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Derg 22(65):481–490
51. Wang S, Jia H, Abualigah L, Liu Q, Zheng R (2021) An improved hybrid aquila optimizer and harris hawks algorithm for solving industrial engineering optimization problems. Processes 9(9):1551
52. Abbasi A, Firouzi B, Sendur P (2021) On the application of Harris hawks optimization (HHO) algorithm to the design of microchannel heat sinks. Eng Comput 37:1409–1428
53. Hu J et al (2022) Detection of COVID-19 severity using blood gas analysis parameters and Harris hawks optimized extreme learning machine. Comput Biol Med 142:105166
54. Ye H et al (2021) Diagnosing coronavirus disease 2019 (COVID-19): Efficient Harris hawks-inspired fuzzy K-nearest neighbor prediction methods. IEEE Access 9:17787–17802
55. Jiang F, Zhu Q, Tian T (2022) Breast cancer detection based on modified Harris hawks optimization and extreme learning

machine embedded with feature weighting. Neural Process Lett 55:1–24. https://doi.org/10.1007/s11063-021-10700-w
56. ŞenolÇelik T, YusufŞengül A, HakanİInci (2018) Investıgatıon of plant and anımal productıon values affectıng consumer prıce ındex by multıvarıate adaptıve regressıon. Splıne: Turkey Case J 3(5):399–408
57. Toprak S (2011) Time series modelling using multivariate adaptive regression splines and conic quadratic programming. Dicle Üniversitesi
58. Zhang W, Goh ATC, Zhang Y (2016) Multivariate adaptive regression splines application for multivariate geotechnical problems with big data. Geotech Geol Eng 34:193–204
59. Al-Sudani ZA, Salih SQ, Yaseen ZM (2019) Development of multivariate adaptive regression spline integrated with differential evolution model for streamflow simulation. J Hydrol 573:1–12
60. Celik S (2019) Comparing predictive performances of tree-based data mining algorithms and MARS algorithm in the prediction of live body weight from body traits in Pakistan goats. Pak J Zool 51(4):1447–1456
61. Özfalcı Y (2008) Multivariate adaptive regression splines: MARS. Gazi Üniversitesi, Ankara
62. Kuter S (2014) Atmospheric correction and image classification on MODIS images by nonparametric regression splines
63. Di W (2006) Long term fixed mortgage rate prediction using multivariate adaptive regression splines. School of Computer Engineering Nanyang Technological University
64. Yerlikaya F (2008) A new contribution to nonlinear robust regression and classification with MARS and its applications to data mining for quality control in manufacturing. Middle East Technical University
65. Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 39(4):561–577
66. Smith BJ, Hillis SL (2022) MATLAB toolbox for ROC analysis of multi-reader multi-case diagnostic imaging studies. İn: Medical ımaging 2022: ımage perception, observer performance, and technology assessment, SPIE, 2022, pp 99–111
67. Gu Q, Cai Z, Zhu L, Huang B (2008) Data mining on imbalanced data sets. İn: 2008 ınternational conference on advanced computer theory and engineering, IEEE, pp 1020–1024
68. Mirzaei B, Nikpour B, Nezamabadi-pour H (2021) CDBH: A clustering and density-based hybrid approach for imbalanced data classification. Expert Syst Appl 164:114035
69. Wong GY, Leung FHF, Ling S-H (2018) A hybrid evolutionary preprocessing method for imbalanced datasets. Inf Sci (NY) 454:161–177
70. Lim P, Goh CK, Tan KC (2016) Evolutionary cluster-based synthetic oversampling ensemble (eco-ensemble) for imbalance learning. IEEE Trans Cybern 47(9):2850–2861
71. Yavuz BÇ, Yildiz TK, Yurtay N, Pamuk Z (2014) Comparison of k nearest neighbours and regression tree classifiers used with clonal selection algorithm to diagnose haematological diseases. AJIT-e Acad J Inf Technol 5(16):7–20
72. Jaiswal M, Srivastava A, Siddiqui TJ (2019) Machine learning algorithms for anemia disease prediction. İn: Recent trends in communication, computing, and electronics: select proceedings of IC3E 2018, Springer, Berlin, pp 463–469