MUŞ ALPARSLAN ÜNİVERSİTESİ

HENDESE TEKNİK BİLİMLER VE MÜHENDİSLİK DERGİSİ

MUŞ ALPARSLAN UNIVERSITY

JOURNAL OF HENDESE TECHNICAL SCIENCES AND ENGINEERING

Muş Alparslan Üniversitesi Külliyesi Teknik Bilimler Meslek Yüksekokulu, 49250, Muş, Türkiye

https://dergipark.org.tr/en/pub/hendese/board

e-ISSN: 3023-7777

HENDESE Ekim 2025, 2 (2): 44-102

Cilt / Volume: 2 **Sayı / Issue: 2** **Yıl / Year: 2025**

Optimizing Anemia Diagnosis in Children and Adolescents Using Support Vector Machines with Feature Selection and Genetic Algorithms

Nuri Korhan[1]

[1] *Istanbul Technical University, Maslak, Istanbul, TÜRKİYE*

nurikorhan@gmail.com —ORCID 0000-0003-4351-2885

Article Information

Article Types	Research Article
Received	03 October 2025
Accepted	27 October 2025

Year: 2025 | **Volume:** 2 | **Issue:** 2 | **Pages:** 84-88

Cite as: N. Korhan, "Optimizing anemia diagnosis in children and adolescents using support vector machines with feature selection and genetic algorithms," *Hendese Journal of Technical Sciences and Engineering*, vol. 2, no. 2, pp. 84–88, 2025, doi: 10.5281/zenodo.17474575.

Corresponding Author: Nuri KORHAN

https://doi.org/10.5281/zenodo.17474575

HENDESE TEKNİK BİLİMLER ve MÜHENDİSLİK DERGİSİ
JOURNAL of HENDESE TECHNICAL SCIENCES and ENGINEERING

RESEARCH ARTICLE

Optimizing Anemia Diagnosis in Children and Adolescents Using Support Vector Machines with Feature Selection and Genetic Algorithms

Nuri KORHAN[*1]

[1]*Istanbul Technical University, Maslak, Istanbul, TÜRKİYE*

(**Received:** 03.10.2025, **Accepted:** 27.10.2025, **Published Online:** 30.10.2025)

Keywords
SVM,
Feature Selection,
Genetic Algorithm,
Fisher Score,
Medical Data Classification

ABSTRACT

Support Vector Machines (SVMs) are widely used learning methods that often achieve remarkable results, encouraging further research into their applications. This paper presents a paradigm based on classification via SVMs for diagnosing anemia in children and adolescents (people under 18 years of age). As training and test data, hemogram test results of 50 individuals (either patients or healthy) are used. Input data consists of five different features (HGB, HCT, MCV, MCH, and MCHC). In order to increase the classifier's efficiency, feature subset selection is applied, and the number of features is decreased. The Fisher score algorithm obtains the most important features for this preprocessing step. These selected features were then used to train the SVM. After repeated training sessions, it has been observed that the performance depends heavily on not only the input's selected feature subsets but also the SVM's hyperparameters. To improve performance (in terms of accuracy), the penalization coefficient of the slack variable is optimized by a well-known optimization method called the genetic algorithm. Experimental results demonstrate that models trained with only two features (HGB and HCT) achieved the highest average accuracy (99.5%), average sensitivity (100%), and average specificity (99%). These findings suggest that feature selection and parameter tuning significantly improve diagnostic performance while reducing computational cost. The proposed framework highlights the potential of hybrid SVM models as decision-support tools for medical diagnostics.

***Corresponding Author:** nurikorhan@gmail.com

Volume: 2 **Issue:** 2

Cite as: N. Korhan, "Optimizing anemia diagnosis in children and adolescents using support vector machines with feature selection and genetic algorithms," *Hendese Journal of Technical Sciences and Engineering*, vol. 2, no. 2, pp. 84–88, 2025, doi: 10.5281/zenodo.17474575.

Doi: 10.5281/zenodo.17474575

1. INTRODUCTION

Support Vector Machines (SVMs) are efficient tools for pattern classification that have been used successfully in medical diagnosis. SVMs have also been combined with various optimization methods such as simulated annealing, genetic algorithm, or Particle Swarm Optimization (PSO). In Hui-Ling Chen *et al.* [1] constructed a model including SVMs, Feature Selection (FS), and Particle Swarm Optimization (PSO) to diagnose thyroid disease. In Kousarrizi *et al.* [2] analyzed the effects of different feature selection methods on SVM classifiers diagnosing thyroid disease. In Javad Salimi Sartakhti *et al.* [3] presented a work on the diagnosis of hepatitis disease via SVMs (for classification) and Simulated Annealing (for optimization of parameters).

Anemia [4] is referred to as a decrease in the amount of hemoglobin in the blood. It can also be referred to as the reduced ability of the erythrocytes to contain hemoglobin.

In both cases, the effect of the disease is a decrease in the amount of oxygen transported by blood. Anemia can be due to decreased red blood cell production, blood loss, or increased red blood cell destruction. There are many factors that provoke anemia, such as trauma, gastrointestinal bleeding, iron deficiency, thalassemia, a lack of vitamin B12, sickle cell anemia, some contagions (e.g., malaria, thalassemia), and so on.

Being the most widespread disorder of the blood, anemia affects about a quarter of people in the world. Iron-deficiency anemia affects nearly 1 billion people worldwide [5].

Anemia is typically diagnosed on a complete blood exam. Doctors use various data, such as HGB, HCT, MCV, MCH, and MCHC obtained from a complete blood count, to diagnose anemia. A complete blood exam also provides information on the type of anemia [6].

Diagnosis of anemia via machine learning techniques has been studied in many works. However, hybrid methods based on SVM have received relatively less attention in the diagnosis of this disease. In [7], the authors presented a model for the diagnosis of anemia in children and adolescents via Artificial Neural Networks. In Volkan Seymen *et al.* [8] investigated methods for diagnosing anemia in children, adolescents, and adults using a Feedforward Backpropagation Neural Network. This paper presents a systematic approach based on SVMs to diagnose anemia, containing FS and a genetic algorithm, based on classification via SVMs for diagnosing anemia in individuals under 18 years old. The dataset that is used to perform this application was obtained from [7], which focused on anemia diagnosis in children and adolescents via Neural Networks.

2. SUPPORT VECTOR MACHINES IN CLASSIFICATION

SVMs are learning algorithms used to examine the data and identify the classes. SVM was initially proposed as a supervised learning model for binary classification problems and has since been extended to support both classification and regression applications [9]. By default, SVM is a binary, linear, and non-probabilistic classifier. To extend SVMs into nonlinear classification tasks, the "Kernel Trick" is applied. Kernel functions project input vectors into a higher-dimensional feature space, where a linear decision surface can be constructed. The generalization ability of the SVMs is achieved through the properties of the decision surface and kernel-based transformations. Fig. 1 depicts how SVMs separate data into classes using margins [9].

In SVMs, the goal is to find a decision boundary that maximizes the width of the margin separating the samples of one class from another class. Given a set of training data points, the purpose is to construct an optimal decision surface that can separate two different classes [10]. This is achieved by formulating a quadratic optimization problem, which is then solved to obtain the separating hyperplane.

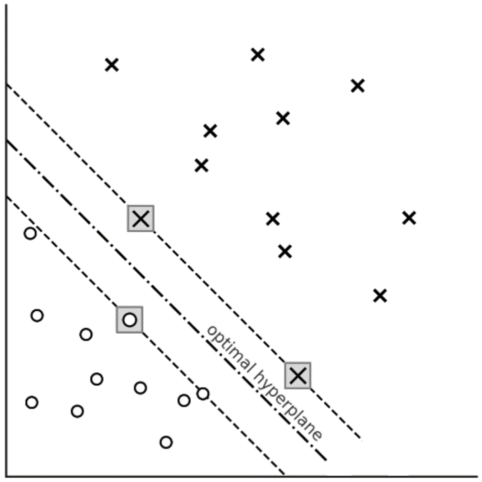

Fig. 1. Support vector machines [9].

A labeled dataset is needed to train an SVM model. Let $\{x_i, y_i\}, i=1,2,\ldots,m$, $y_i \in \{-1,1\}$, $x_i \in R^d$, where x_i represents input feature vectors, and y_i denotes the class labels of corresponding input vectors. The separating surface can be formulated as seen in Eq. (1).

$$f(x) = w^T x + b \tag{1}$$

Where w is the weight vector perpendicular to the separating surface and b is the bias. Also, $b/\|w\|$ will be the distance from the origin to the surface. By solving the following optimization problem, the SVM determines the maximum (or in some cases, optimal margin - if not separable) margin of the hyperplane.

$$\text{minimize } \|w^2\| / 2 + C \sum_{i=1}^{\infty} \xi_i \tag{2}$$

$$\text{Subject to: } \quad \xi_i \geq 0, y_i(\langle w, x_i \rangle + b) \geq 1$$

where C is a penalization constant and ξ_i represents the slack variable.

$$\max W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i, \alpha_j, y_i, y_j, x_i^T, x_j \tag{3}$$

Subject to: $\sum_{j=1}^{m} \alpha_i, y_i = 0$ and $\alpha_i \geq 0$

By applying the Karush-Kuhn-Tucker (KKT) conditions, the Lagrangian dual form with multipliers α_is are obtained in this step. Then, bias (b) and weight vector are computed based on the values of multipliers α_i. At the maximum margin, w is calculated as seen in Eq. (4).

$$w = \sum_{j=1}^{m} \alpha_i, y_i, x_i \tag{4}$$

The bias term b can then be calculated based on Eq. (5). The linear decision function can be determined as seen in Eq. (6).

$$\alpha_i[y_i(\langle w, x_i \rangle + b) - 1] = 0, i = 1,2, \dots m \tag{5}$$

$$F(x) = \mathrm{sgn}(\sum_{i=1}^{m} y_i, \alpha_i \langle x, x_i \rangle + b) \tag{6}$$

A linear hyperplane classifier works well in linearly separable cases. However, in nonlinear cases, the input space must be mapped to a higher dimensional feature space where pattern can be separated by a linear classifier. Various kernel functions can be used for this mapping, including quadratic, Gaussian, and polynomial kernels. Based on several kernel experiments conducted in this study, a polynomial kernel is selected. This kernel is defined as seen in Eq. (7).

$$K(x, x_i) = \left(1 + x_j^T, x_i\right)^p \tag{7}$$

where p is the order of the kernel. (In this study, a third-order kernel is used.) After constructing the hyperplane in Equation (1) and identifying the support vectors, the two classes can then be separated by maximizing W in Eq. (8).

$$\max W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i, \alpha_j, y_i, y_j, K(x_i, x_j) \tag{8}$$

subject to: $\sum_{j=1}^{m} \alpha_i, y_i = 0$.

where $\alpha_i \geq 0$ for all i=1,2, 3, …, m and therefore, α_is are calculated by solving the quadratic equation and w is determined depending on α_is. Applying the kernel trick, the decision function can be described as seen in Eq. (9).

$$F(x) = \mathrm{sgn}(\sum_{i=1}^{m} y_i, \alpha_i K(x, x_i) + b) \tag{9}$$

3. FEATURE SELECTION VIA FISHER SCORE

The Fisher score is a feature selection method that identifies the most important features for classification [1]. It is used to reduce the number of dimensions in the feature space, thereby creating a feature space that is easier to train [11]. For each attribute, a Fisher score is calculated by employing Fisher's criterion. Subsets having higher scores are considered to be more important in the decision-making process [1]. The primary purpose of applying Fisher Score is to select features that enable easier class separation. Ideally, features are desired for which the variance of the data distribution within each class is small, while the mean values of the classes are as distant from each other as possible [12].

In this paper, the Fisher Score is determined by employing a criterion function as described in [12]. In the experiments, both highly informative and less relevant features were obtained, as summarized in Table I.

This intuition is expressed by Fisher's criterion:

$$F(J) = \frac{\sum_{c=1}^{C} n_c \left(\mu_c{}^{(j)} - \mu^{(j)}\right)^2}{\sum_{c=}^{C} n_c \sigma_c^{(j)2}} \tag{10}$$

where $F(J)$ denotes the Fisher Score of the j^{th} feature, n_c is the number of samples in class *c*, $\mu_c{}^{(j)}$ is the mean of feature *j* within class *c*, $\mu^{(j)}$ is the overall mean of feature *j*, and $\mu_c{}^{(j)2}$ is the variance of feature *j* within class *c*. Features with higher Fisher Scores provide stronger discriminative power and are therefore preferred for classification tasks. In this work, only the highest-ranked features based on the Fisher Score were selected for subsequent training, allowing the classifier to concentrate on the most informative attributes.

4. OPTIMIZATION OF ANEMIA DIAGNOSIS USING SUPPORT VECTOR MACHINES

In this paper, a classification system based on SVMs was developed to diagnose anemia in children and adolescents. Hemogram test results from 50 individuals were used for both the training and test phases (60% random training samples, 40% held out for test). The feature space was defined by five attributes: hemoglobin (HGB), hematocrit (HCT), mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), and mean corpuscular hemoglobin concentration (MCHC). The output class labels were assigned in binary form, with "1" representing anemic (diseased) and "0" representing healthy. Feature Selection was performed using the Fisher score algorithm to improve classification accuracy. Three experiments were conducted: One using the original dataset and two using reduced datasets constructed from the most important features identified by Fisher Score.

During the experiments, it was observed that the classifier's performance depended not only on the selected feature subsets but also on the parameter settings of the SVM. In particular, the penalization constant C (the soft margin parameter controlling the slack variable penalty) strongly influenced classification accuracy. To optimize this parameter, a genetic algorithm was employed to search for the value of C that maximized the average accuracy of the classifier.

The evaluation metrics are defined as follows:

Accuracy (Acc): Ratio of correctly classified cases (positive or negative) to the total number of cases.

$$\mathrm{Acc} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{11}$$

Sensitivity (Sens): Ratio of correctly classified positive cases to the total number of actual positive cases.

$$\mathrm{Sens} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{12}$$

Specificity (Spec): Ratio of correctly classified negative cases to the total number of actual negative cases.

$$\mathrm{Spec} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{13}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. To account for variability across multiple runs, the average accuracy is computed as seen in Eq. (14).

$$\mathrm{AveAcc} = \frac{\sum_{i=1}^{m} \mathrm{Acc(i)}}{\mathrm{m}} \tag{14}$$

where Acc (i) denotes the ith iteration, and m is the total number of iterations.

5. RESULTS

In this study three different experiments are carried out by utilizing three different datasets. The first dataset included all features of the original dataset. The second dataset includes the top three features (HGB, HCT, MCV), and the third dataset includes the top two features (HGB, HCT). The Fisher scores of the feature subsets are presented in Table I. In the optimization part of the experiments, the optimal value of the coefficient C was calculated by genetic algorithm and selected as the soft margin parameter of the classifier. Based on this optimized parameter, performance metrics of the classifier were then evaluated.

Accuracy-wise comparative analysis: The highest average accuracy of 99.5% was achieved when only the top two features (HGB and HCT) were utilized. The second-best performance, with an average accuracy of 94.5%, was obtained using the top three features (HGB, HCT, and MCV). Finally, when all features were employed, the model attained an average accuracy of 93%.

Table I. Fisher scores of the feature subsets.

Feature	Score
HGB	3.0684
HCT	2.8386
MCV	0.0623
MCH	0.0599
MCHC	0.0083

Sensitivity-wise comparative analysis: In terms of sensitivity, the best results were obtained with the top two features, achieving an average sensitivity of 99%. The top three features yielded 95% while all features yielded 92%.

Specificity-wise comparative analysis: In terms of specificity, the best results were again obtained with the top two features, achieving an average Specificity of 100%. Top three features yielded 95% while all features also yielded 95%.

The results of the experiment using all five features (HGB, HCT, MCV, MCH, MCHC) are presented in Table II. The optimal value of C was 0.0888, determined by genetic algorithm. Maximum accuracy, sensitivity, and Specificity of 100% were observed. On average, the model obtained 93% accuracy, 91% sensitivity, and 95% Specificity, reflecting strong but slightly less consistent results compared to models trained with fewer features.

Table II. Experiment results for the feature subsets (HGB, HCT MCV, MCH, MCHC).

Test no	Acc%	Sens%	Spec%
1	95	100	90
2	100	100	100
3	95	90	100
4	95	90	100
5	95	90	100
6	85	90	80
7	85	80	90
8	95	90	100
9	100	100	100
10	85	80	90

The results of the experiment using HGB, HCT, and MCV are shown in Table III. Optimal value of C for that experiment was determined to be 0.1834. The model achieved maximum values of 100% for all metrics. On average, it achieved 94.5% accuracy, 95% sensitivity, and 95% Specificity, indicating reliable classification performance across all measures.

Table III. Experiment results for the feature subsets (HGB, HCT, MCV).

Test no	Acc%	Sens%	Spec%
1	90	90	90
2	100	100	100
3	95	100	90
4	100	100	100
5	90	90	90
6	95	90	100
7	95	90	100
8	95	90	100
9	95	100	90
10	90	100	90

The results of the experiment using features HGB and HCT are shown in Table IV. The optimal value of C was found to be 0.2252 using the genetic algorithm. The model achieved maximum values of 100% in accuracy, sensitivity, and Specificity. On average, it reached 99.5% accuracy, 99% sensitivity, and 100% specificity, demonstrating that while perfect performance was attained in certain iterations, the overall performance remained consistently strong.

Table IV. Experiment results for the feature subsets (HGB, HCT).

Test no	Acc%	Sens%	Spec%
1	100	100	100
2	100	100	100
3	100	100	100
4	100	100	100
5	100	100	100
6	100	100	100
7	100	100	100
8	95	90	100
9	100	100	100
10	100	100	100

6. CONCLUSION AND FUTURE WORK

In this study, a system was developed to diagnose anemia in children and adolescents using SVMs. Improved accuracy was achieved by selecting the most relevant features and determining the optimal classifier parameter value for each dataset based on feature importance. The results demonstrated that both feature selection and parameter optimization significantly enhanced when only two feature subsets were used, which proved more efficient in terms of both computational cost and classification accuracy. Future studies will expand validation cohorts with larger datasets and explore broader clinical applications.

ACKNOWLEDGEMENTS

Acknowledgements are given to the authors of [7] for providing access to the anemia dataset used in this work.

REFERENCES

[1] H. L. Chen, B. Yang, G. Wang, J. Liu, and D. Li, “A three-stage expert system based on support vector machines for thyroid disease diagnosis,” *Journal of Medical Systems*, vol. 36, pp. 1953–1963, Jun. 2012, doi: 10.1007/s10916-011-9655-8.

[2] M. N. Kousarrizi, F. Seiti, and M. Teshnehlab, “An experimental comparative study on thyroid disease diagnosis based on feature subset selection and classification,” *International Journal of Electrical & Computer Sciences (IJECS–IJENS)*, vol. 12, no. 1, pp. 13–20, 2012.

[3] J. S. Sartakhti, M. H. Zangooei, and K. Mozafari, “Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA),” *Computer Methods and Programs in Biomedicine*, vol. 108, no. 2, pp. 570–579, Nov. 2012, doi: 10.1016/j.cmpb.2011.08.003.

[4] A. V. Hoffbrand, P. A. H. Chowdhry, G. R. Collins, and J. Loke, *Hoffbrand’s Essential Haematology*, 9th ed. Hoboken, NJ, USA: Wiley-Blackwell, 2024.

[5] A. Baldi and S. R. Pasricha, “Anaemia: Worldwide prevalence and progress in reduction,” in *Nutritional Anemia*, C. D. Karakochuk, M. B. Zimmermann, D. Moretti, and K. Kraemer, Eds. Cham, Switzerland: Springer, 2022, pp. 3–17, doi: 10.1007/978-3-031-14521-6_1.

[6] B. F. Rodak, G. A. Fritsma, and E. M. Keohane, *Hematology: Clinical Principles and Applications*, 5th ed. St. Louis, MO, USA: Elsevier, 2016, pp. 150–172.

[7] A. T. Koru, E. Akdoğan, E. Kaya, M. Aktan, and A. Koru, “Diagnosis of anemia in children via artificial neural network,” *International Journal of Intelligent Systems and Applications in Engineering*, vol. 3, no. 1, pp. 24–27, Jan. 2015, doi: 10.18201/ijisae.61036.

[8] S. Volkan, G. D. Çelik, and A. Devrim, “The diagnosis of iron-deficiency anemia using feedforward backpropagation neural network,” *Journal of Medical and Technological Innovations*, vol. 2, no. 1, pp. –, 2014.

[9] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[10] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2nd ed. New York, NY, USA: John Wiley & Sons, 2001, pp. 207–212.

[11] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, Sep. 1936, doi: 10.1111/j.1469-1809.1936.tb02137.x.

[12] Q. Gu, Z. Li, and J. Han, “Generalized Fisher score for feature selection,” *arXiv preprint* arXiv:1202.3725, Feb. 2012, doi: 10.48550/arXiv.1202.3725.