Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2025.00038

Research Article

Determining the parameters of data mining techniques and their effects on iron deficiency based prediction

Sahar J. MOHAMMED[1], Ahmed Kh. ABBAS[1], Arshed A. AHMAD[1], Mohammed S. MOHAMMED[1], Murat SARI[2], Hande USLU TUNA[3,*]

[1]*University of Diyala, College of Education for Pure Sciences, Diyala, 32001, Iraq*
[2]*Department of Mathematical Engineering, Istanbul Technical University, Istanbul, 34469, Türkiye*
[3]*Department of Mathematics, Yildiz Technical University, Istanbul, 34220, Türkiye*

ARTICLE INFO

Article history
Received: 14 February 2024
Revised: 09 April 2024
Accepted: 11 May 2024

Keywords:
Anemia; Attribute Selection; Prediction; Sequential Minimal Optimization; Local Weighted Learning

ABSTRACT

Training datasets are not the only elements affecting the overall prediction system; data mining parameters also have effects on the implementation processes that need to be taken into account. The purpose of this research is to investigate the influence of the main characteristics of the most used data mining approaches on anemia prediction. In this context, for the K-Nearest Neighbour (K-NN) approach, it is critical to define the k-value to specify the number of points used to measure the distance between various types of classes. Furthermore, the Local Weighted Learning (LWL) has a kernel value that specifies the width of the search process used to generate the LWL weight function. The Sequential Minimal Optimization (SMO) has an n-tuple alpha value that is determined by the training data in order to meet the Kraush Kuhh Tucker (KKT) condition and speed up the prediction process. When a superior choice is optimized for each strategy, these data mining methods are shown to produce high-performance predictions. It has also been noticed that the number of features and dataset size have an impact on the performance of these methods. In this study, feature selection methods and mining methods are compared in terms of appropriate selection of parameters and dependency on dataset information. The methods proposed here have predicted anemia more accurately than prior versions of each method. For the applied dataset, the features are reduced from 11 to 8. In addition to this feature reduction and parameter selections of a good method, i.e. K-NN, has an increase of about 3.8% in prediction performances based on the proposed model.

Cite this article as: Mohammed SJ, Abbas AK, Ahmad AA, Mohammed MS, Sarı M, Uslu Tuna H. Determining the parameters of data mining techniques and their effects on iron deficiency based prediction. Sigma J Eng Nat Sci 2025;43(2):655–664.

INTRODUCTION

Various studies reveal that anemia is the most known blood disease in the world [1]. According to the World Health Organization (WHO), anemia is a disease in which the number of red blood cells and therefore the oxygen-carrying limit are not sufficient to meet the needs of the body [2]. Ordinary hemoglobin and hematocrit levels vary with

*Corresponding author.
*E-mail address: usluh@yildiz.edu.tr

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

age and gender. If the value falls below the normal adequacy limit for the age and gender, frailty exists. The researchers led this study to assess 189 nations, both sexes, and 20 distinct age groups, using information and resources from an existing dataset uploaded by WHO in 2010, and to investigate the severity of the condition worldwide. They discovered that 32.9% of people globally suffer from iron deficiency. Iron deficiency is more common in children under the age of five and women. The most prevalent kind of weakness is iron deficiency pallor [3]. Pallor, which largely impacts personal pleasure, is both an illness and a side effect of many real infections, so treating it can be critical in general, and obtaining the correct result is the first step toward therapy. Anemia affects roughly 614 million women and 280 million children, according to practical and statistical estimates [4]. When AdaBoost was utilized, that study demonstrated excellent performance in terms of time, additional assessment procedure, and Orange tool performance. That study emphasizes the possibility of utilizing machine learning to enhance health decisions for women and children.

Various machine learning techniques were also used for this purpose [5], with the researchers relying on demographic data, specifically non-Hispanic black women and naturalized citizens. According to the researchers, DT fared best in their trials and can be utilized to improve healthcare by delivering low-cost, high-quality care. Some other studies, such as [6], focused solely on children and those under the age of 5. Their studies concentrated solely on blood tests and other anemia-related indicators. The researchers used a Random Forest Tree to apply their technique to a noisy dataset. These results are improved by approximately 0.2% for a balanced dataset and the same RFT version. According to the rule-based paradigm reported in the literature [7], researchers and health organizations generally included women, especially pregnant women, in their studies. Researchers identify the key features that influence prediction performance, such as age, nutritional education status, and diversity. The researchers employed OneR for this purpose and improved the prediction accuracy for pregnant women by roughly 12.4%. Furthermore, novel hybrid techniques have been developed to forecast this disease based on pertinent biological data [8]. In that study, an enhancement for this goal was proposed, using Binary PSO as a feature selection method. The PSO was used in combination with the SVM approach to produce a reliable treatment decision.

In addition, the effect of the number of patients and non-patients was investigated in the literature as a major determinant [9]. To accomplish this, image augmentation techniques were utilized on anemia-related medical images from selected samples, which were then combined with Colour space mapping for data analysis using the ANN. In the same prediction procedure, an Extreme type machine learning application was proposed to improve various strategies, such as a single layer Feed Forward network version in the hidden section [10]. This model identified beta thalassemia phenotype and iron deficiency. In a study in North East India [11], five ML approaches, including Naïve Bayes, were successfully utilized. However, only Random Forest outperformed the other approaches with 15 characteristics. In addition, many datasets encompassing around ten thousand patients with various kinds of anemia were reviewed [12]. NB and other techniques were used to detect these types at a minimal cost and in a short amount of time with good performance. Because of the diversity and applicability of the anemic dataset, the researchers decided to include other relevant patients, such as young female students [13]. In a study on anemia prediction based on palm-to-spot differences, five techniques were applied [14]. In the corresponding study, NB provided higher performance for different datasets collected from various hospitals. It is seen that these studies can play a role in guiding many researchers [15]. In the relevant study, researchers claimed that 40% of children under five years of age are anemic. In that paper, a feature selection with RFT was applied to provide accurate results and indicate the relationship between variables. Also, factors affecting treatment decisions such as drug availability and legal restrictions are indicated in the study. Computer-aided decision making and analysis are widely used in the medical industry. In this study, a technique was devised to help and support the specialist in detecting different types of anemia. In the same context, previous work on predicting desired types was examined and assessed [16, 17]. Hybrid models have also been utilized in the literature [18-20]. A computer-aided system was developed to present research in medical education. This was one of the first studies to use a computer to diagnose anemia. PlanAlyzer [21], introduced to the academic community for a diagnostic process such as heart disease, was designed to clarify approaches for students to identify a common medical condition. Furthermore, researchers reported in one study that after testing and assessment, the curriculum was reviewed and introduced to cardiology and hematology departments to teach the diagnosis of a condition such as anemia and chest discomfort. The WEKA data mining tool has been also used to design a classification system to identify types of anemia based on multiple parameters extracted from the samples. The researchers applied various techniques to detect four types of anemia with 10 attributes. The C4.5 decision tree method had a success rate of 99.42%, which was higher than the 88.13% success rate of support vector machines [22].

Here, the effectiveness of various approaches such as data mining in anemia prediction has been investigated comparatively by considering the primary parameters. The K-value, which indicates the number of points to measure the distance to various class types, should be defined for the K-NN technique. In addition, a kernel value (T) for the LWL indicates the search width used to determine the LWL weight function. The KKT condition is satisfied by the SVM with n-tuple alpha values dependent on the training data to speed up the prediction process. This study analyzes feature selection and mining strategies based on appropriate parameter selection and dataset. Here, we consider the

applicable dataset, which consists of eight features. Except for feature reduction and appropriate parameter selection, the K-NN improves prediction performance by 3.8%.

DATASET

Samples have taken for 539 patients with 11 features. In this study, the features for each subject were also reduced to 8. The samples include blood variables which have been read for each subject such that, Hemoglobin (HB), Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin Concentration (MCHC), White Blood Cell (WBC), Platelets (PLT), Red Blood Cells (RBC) and sex and age that were reported in the literature [23, 24]. The Hemoglobin (HB) located within the RBCs is a transportable protein that is composed of iron atoms. The RBCs are concave cells, and their nuclei contain the hemoglobin, but the nuclei are not useful. The MCH is an estimation that is determined from the HB level and the quantity of RBC. WBCs are responsible for safeguarding the body from infectious illnesses. HCT is a measure of the proportion of red blood cells in each volume of blood. MCHC is the amount of the concentration of blood in each area. PLTs are small, disc-shaped elements in the blood that help in clotting and are also classified as blood cells. MCV is the same expression of MCHC but for a specific sample, and other biophysical variables like gender and age are also considered. Because the natural hemoglobin levels in the body differ between males and females, the ratio is typically one to two, with males having higher levels, and vary according to age. For the data, it has been considered that blood diseases are (1) iron-deficiency anemia, (2) deficiency in vitamin B12, (3) thalassemia, (4) sickle cell, and (5) spherocytosis.

CLASSIFIERS

Classification is the procedure of predicting to which class a given set of data points belongs to. For this study, it is necessary to design an approximation of a mapping function (f) that transforms the input data (X) into the desired discrete values (y), considered as a prediction system. Data science utilizes classifiers, which are a kind of machine learning algorithm, to designate a class identification to data input. Classifier algorithms use advanced mathematical and statistical techniques to calculate the probability of input being categorized in a certain manner. In this study, the following classifiers have been utilized. For defining parameters that are related to functions, such as ML, [25] presented a minimal defining of blocks in shift space. This study established the properties of synchronized components in sub-shifts which are dependent on many classifiers such as two applied classifiers in this paper.

K-Nearest Neighbor (K-NN)

K-NN is a simple algorithm to seek a new point that belongs to different classes based on an equal measurement. The term neighbors means that any point that is considered a new point must belong to an old neighboring point by adding it (generic old instances). Suppose there is more than one point, then by adding them it will be classified according to the closest class. According to Figure 1, K-NN shows that the prediction varies with the value of k. For example, if k is equals to 10, this means that 10 points are in the classification circle and these points have 4 classes. K-NN will assign new points to any close and counted class. As shown in Figure 1, there are newly added points to the samples (2, 4, 3 and 1 point) belonging to each class (Disease type 1, 2, 3 and type 4) respectively. Starting with 10 as k-value means selecting 10 points close to a center point, according to Figure 1, the majority of points belong to disease type 2 (4 points), therefore, K-NN can be written as a class 2 for k is equals to 10. When k is equals to 13, a new point is added to the old instances so that there are 19 points. But at this value of k, K-NN is represented as a disease type 4 due to the newly added points. In this paper, it was important to choose the value of k to give optimal solutions using the cross-validation technique. To find the minimum distance between points, it is done by applying two types of distance calculation methods (Euclidean and Manhattan). K-NN can be applied for disease prediction, such as heart disease as in [26], or it can be combined with a genetic algorithm as in [26], with similar types of measurement factors. Several papers have proven

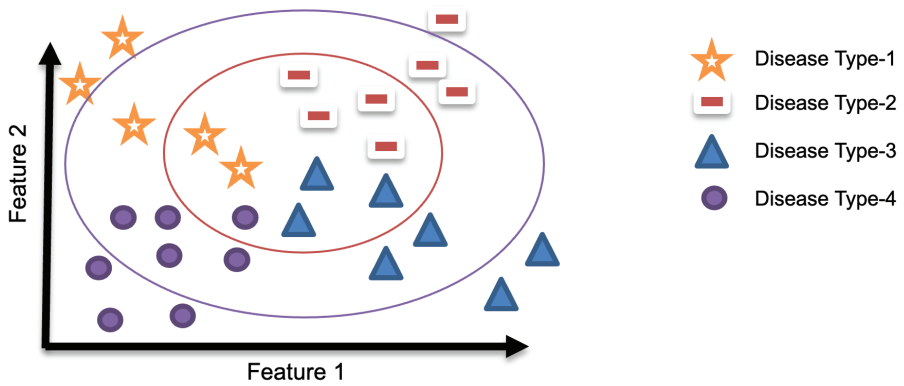


Figure 1. Description of the K-NN algorithm based on K-value.

that K-NN gives better results when compared to other machine learning techniques, as in [27-29].

Locally Weighted Learning (LWL)

LWL is not a parametric method, it depends on the training data. If the shape or pattern of the collected dataset is unknown, the closest points to be tested can be estimated based on the closest samples. The overall prediction process is enhanced by assigning greater weights to data points closer to the tested point. The simple description of LWL in Figure 2 shows how the relation between patient parameters and the disease type of Anemia have measured.

In the LWL, training data is important as a definition of parameter to make a prediction, because the system should specify which points are new to the testing points. In this model, T is the affected parameter to determine the width of the kernel function where the $w(i)$ is the weighted function which measures distances between the tested point and all other points in the training data to provide a weight as in Equation 1.

$$Kernel\ Weight(i) = exp\{-((x^{(i)}-x))^2/2T^2\} \quad (1)$$

Here i is the number of training data, and x is the test point, it showed the effect of the T value on the weight and distance calculation process as shown in Figure 2, when $T = T_1$, LWL has more points close to the tested one. T_2 provided fewer points to the LWL model. The T parameter is chosen according to the training and testing data not defined or even learned by LWL, for this reason, it's called hyperparameter. The denominator of the kernel weight defines the distance condition between the training and test data, if the distance is close to the training data, the denominator is equal to zero, while the weight kernel is usually close to 1 or the highest value of the kernel. If the distance is large, the denominator is larger, then $w(i)$ is close to zero. The differences between the standard and LWL regression mode are shown in Equation 2 to find the optimum value as in [30-32]. Here x represents the training data, w represents the weight, and Y represents the class type.

$$\text{Optimal value} = (x^T.w.x)^{-1}.x^T.w.Y \quad (2)$$

Ripper

It refers to minimizing errors by incremental repeated methods. Ripper modifies the performances of the Decision Tree (DT) by evolving multiple iterations which

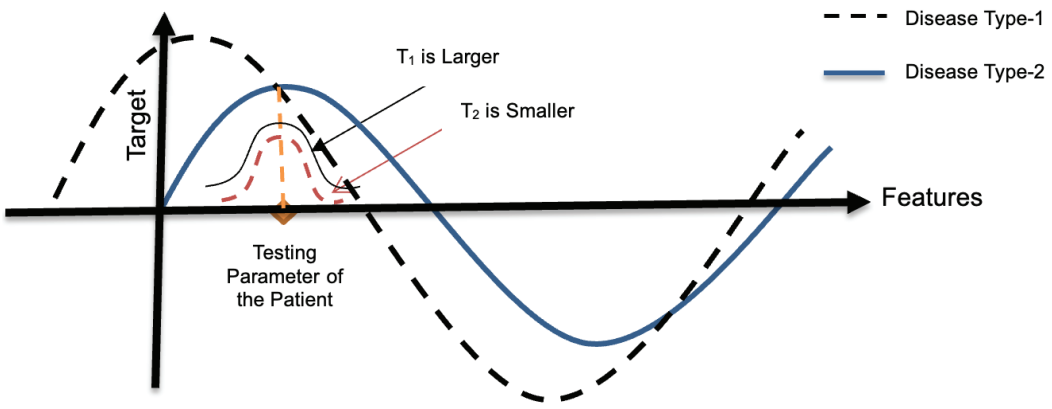


Figure 2. Description of the LWL algorithm based on T-value.

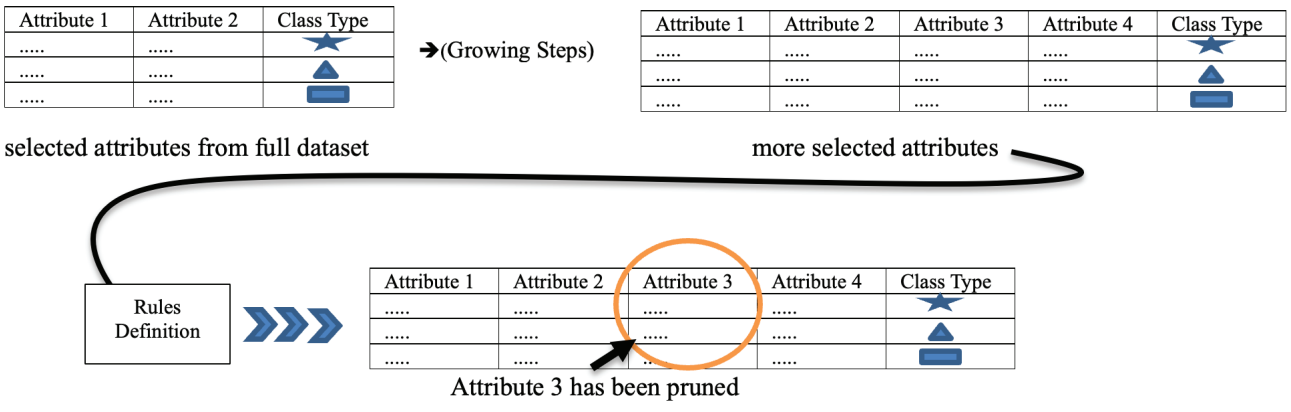


Figure 3. Ripper steps to gain optimal solutions with several epoch numbers.

can be presented in three steps: Growing, Pruning, and Optimization. At the first step, it is applied the same steps of the DT to make the corrected path according to data and its attributes. All attributes are added and checked by the Growing DT process until no longer a need for any other entropy or adding process, then these rules are pruned directly. These steps will be repeated until getting or optimizing the optimal solutions. Ripper applied for disease prediction and compared to other methods such in [33-36]. Figure 3 shows the steps of modifying the steps by using this model.

Sequential Minimal Optimization (SMO)

It is a sequential process to optimize the smallest sub-functions for each iteration step. In this technique, alpha values have been supplied from SMO to satisfy the constraint of the required problem. SMO alpha values are denoted as Lagrange multipliers which can be calculated easily. These multipliers should be identified first before working or applying any SMO Steps. Depending on training data numbers, n values of SMO Lagrange will be selected from all these defined alpha values. This is considered the smallest sub-problem. For any training dataset $n(n\text{-}1)$ possibilities are derived, but at each iteration, two possible values have been selected to accelerate the convergence. By applying the KKT condition to discover the violation samples that need to be optimized in the next iterations. After finding a violation of training samples according to α_i and α_j, these two alpha values will be re-optimized while keeping the rest of the alpha constants. The previous steps are repeated until all alpha values satisfy the KKT Conditions and no further sample violations are found. Any large problem is split into a small subproblem with two alpha values to speed up the computational process as described in [37-39]. Figure 4 shows the anemia prediction steps based on the anemia dataset and considering the backward version of SVM (SMO).

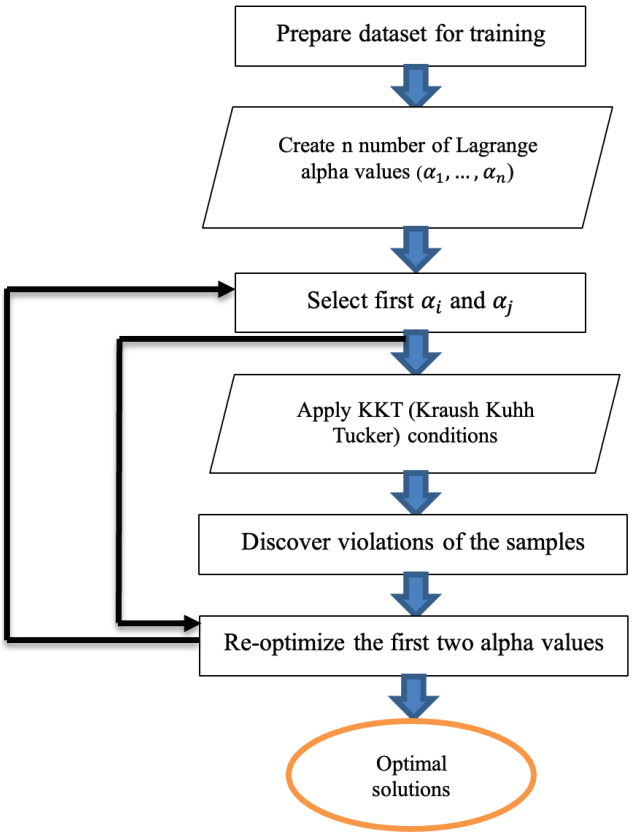


Figure 4. Anemia prediction steps based on the SMO with parameter optimization.

Instance-Based Classifier (K*)

It is known as the Lazy learning process due to simple prediction based on new instance similarity with observed training dataset instead of building a model and following steps. It differs from DT and K-NN in memorizing all samples instead of building any model stored in memory and comparing new samples with saved samples. The decision is based on K* most similar instances, not distance metrics like in K-NN. This approach is considered memory intensive due to storing all datasets with cost increment while dealing with large datasets. It struggles with generalization, especially when training data is noisy or has irrelevant features. It also has fewer parameters to be tuned for the overall process and widely used when having complex relationships between features and classes which makes other techniques difficult to build. In addition, K* is sensitive to outliers, high dimensional features and imbalanced classes when one class type is more repeated than another as mentioned in [40].

Logistic Regression Model (LLRM)

This approach which uses a function known as logistic function to model the relation for a given data belonging to a certain class has been widely used for binary classification tasks. It is also simple to understand and interpret, making it suitable for explaining the relationship between target variables. It is also efficient for small datasets as well as for linearly separable attributes. However, this method is not suitable for complex datasets due to the linear relation assumption by this method. It is also like K* sensitive to outliers but limited to binary classifications which assume that observations are independent of each other leading to mistakes in some datasets as explained more in [41].

Feature Selection

Each data point on the features represents the prediction process; a larger number of features may cause a harder prediction. Minimization of these attributes is important to select the features that have the most impact on the prediction system by providing minimization on prediction time as mentioned in the literature [38]. High measurement information makes it difficult to test and prepare common classification strategies. Determination of the property procedure solves difficulties which is done by a

few strategies like: Attributes Evaluation, Correlation, Pick up Ratio and Principal Component. This is also realized by applying reasonable-looking strategies related to each of the choice processes such as Best First and Ranker. From all the over-selecting strategies, WBC, Gender, and Age are the least affected features on the overall data. Even the feature selection techniques are different, as described in research [39], which provides the impact features of the Diabetic India database. In some articles, further selection techniques have been applied by mixing the Genetic Algorithm (GA) with it as pointed out in the literature [42] or even applying it for different fields not just mining algorithms such in agriculture [43] and biomedical [44].

RESULTS AND DISCUSSION

Samples have been taken for 539 patients with 11 features for the first time of classification and before minimization of the features. The Instance Based Classifier (K*), LWL, RIPPER, SMO, K-NN classifier and Logistic Regression Model (LLRM) techniques have been applied. Each of the techniques has been measured for some parameters like Overall System Accuracy (corrected classified samples), MAE, RMSE as well as each class type parameters such as True Positive (TP) Rate, False Positive (FP) Rate, Precision, and Precision-Recall Curve (PRC) Area. First step was studying all these techniques under the same conditions for 11 attributes and 10-fold validation before utilizing of the feature selection. Table 1 shows the overall accuracy according to 11 attributes for the techniques of interest. This table shows that the SMO has the greatest accuracy among all applied techniques. This is due to kernel transformation which assists to linearization of data. While LWL is the worst classification technique due to the weighted measurement of each sample according to the required neighbor sample which is sort of approximate value. For this reason, firstly data should be clearly identified and minimized for these types of techniques to provide more accuracy to Anemia data classification.

In addition, the same table has shown the results according to two parameters MAE and RMSE. It has shown that SMO had the highest MAE and RMSE related to the same data which proved the benefits of using minimization of the features to make data more linear before dealing with these types of techniques. According to the SMO, weighted samples will be more accurate as distance with well-separated data is indeed classified. MAE refers to the error between each paired sample on the same data, which should be carefully minimized or regularized to ensure that the data is well classified with minimum absolute error. RMSE refers to the differences between the desired prediction samples and the actual corrected samples. This leads to some data parameters being far from the corrected and classified samples. The aim of this work is to reduce these values as much as possible to get the best technique performances for such a type of non-linear data. After applying the feature selection to detect the most affected features on the overall samples as well as performance of the technique. Increasing system prediction efficiency and specifying a suitable method to determine these features so, workers in health institute, hospitals or even programmer can reduce time and cost for such a type of data classification. According to the attribute selection process, this is done by some methods such as Attribute Evaluation, Correlation, Gain Ratio and Principal Component. This is done by applying appropriate search methods related to each selection process: Best First and Ranking. From all the above selecting methods, WBC, Gender and Age were found as the least impact features on the overall data. This might be due to large distances between such a type of attributes could make the overall classification technique worse and less efficient. A new calculation of the same data but for 8 attributes were calculated to be compared with the previous results that related to 11 attributes. Table 2 shows the same calculated parameters for 8 attributes.

The results shown in Figure 5 proved that after deleting three of these unusual attributes made the overall system performances better and enhanced the classification process. Also, applied techniques reached the best algorithm, the SMO, in data classification after omitting three number of attributes. Figure 6 shows the overall MAE and RMSE for the data used. The same results are also obtained.

Another study has been done for this work to prove system enhancement after selection of the most significant data. Figure 7 shows the performance of some classes according to the parameters compared such as TP Rate, FP Rate, Precision and PRC. This has led this study to the

Table 1. Overall system accuracy, MAE and RMSE before feature selection

Method	Accuracy (%)	MAE	RMSE
K*	83.3024	0.0585	0.2202
LWL	76.9944	0.105	0.2311
RIPPER	84.2301	0.0761	0.2154
SMO	84.6011	0.2291	0.3211
K-NN	81.4471	0.0645	0.2472
LLRM	83.7143	0.0758	0.1884

Table 2. Overall system accuracy, MAE and RMSE after feature selection

Method	Accuracy (%)	MAE	RMSE
K*	84.7866	0.0531	0.2119
LWL	76.8089	0.1079	0.2324
RIPPER	85.5288	0.0464	0.2048
SMO	86.6419	0.2286	0.3202
K-NN	84.9722	0.0529	0.2225
LLRM	86.0853	0.0714	0.1849

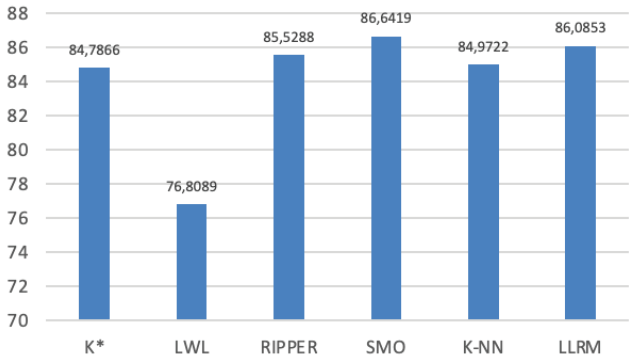


Figure 5. Overall system accuracy for the applied techniques.

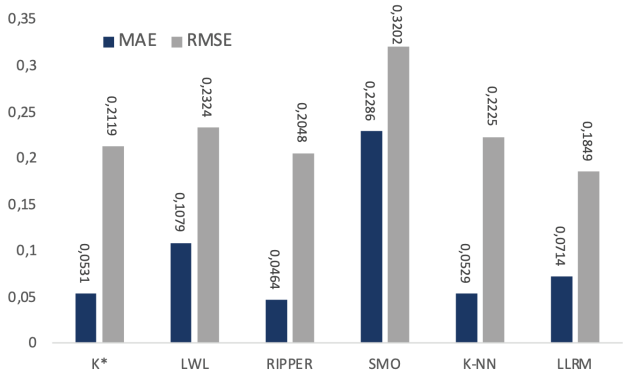
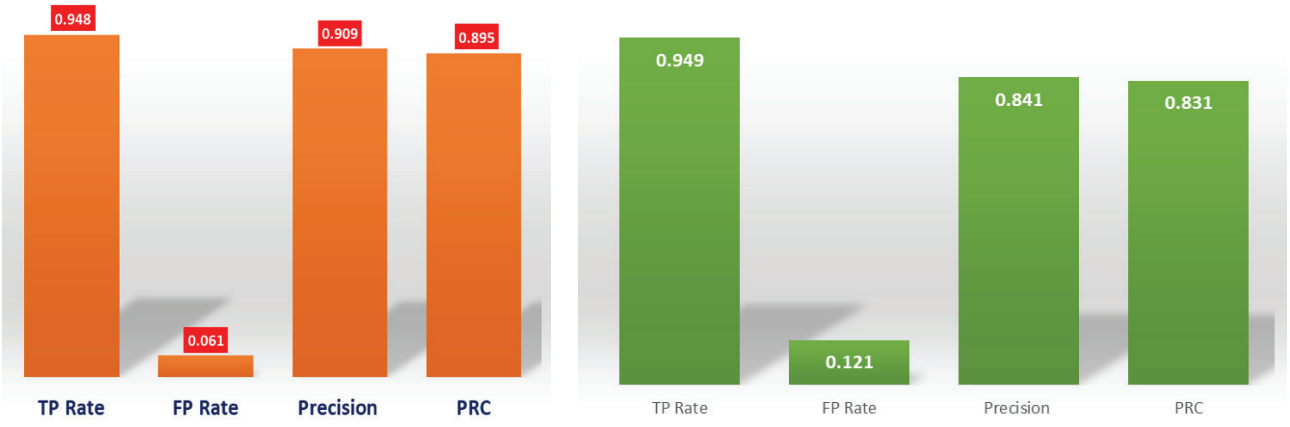


Figure 6. Overall system MAE and RMSE values for the applied techniques.

main points that really need to be identified and specified in these data.

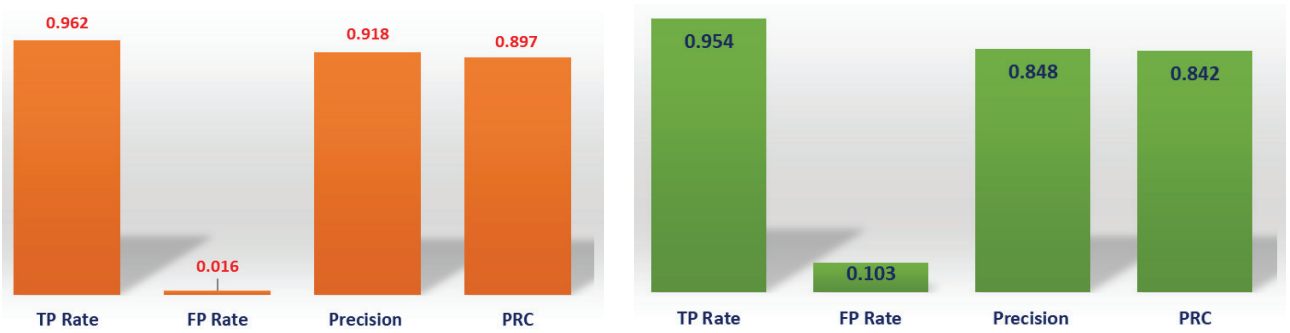
Figure 8 has shown the SMO parameters for the first and third types of anemia after selection of features and deletion of three of the unusual features and shown that the SMO performances change after deletion of some of the non-essential features to obtain better defined samples than before. This is due to the data mining techniques depend on the distance of the data.

The results have been collected for all anemia classes and for all cases before and after feature selection. This study has also been applied for all the mentined techniques



(a) (b)

Figure 7. SMO parameters for (a) the first type and (b) the third type of anemia without feature selection.



(a) (b)

Figure 8. SMO parameters for (a) the first type and (b) the third type of anemia with feature selection.

to give a brief information about the impact of features on data inspection techniques. All techniques have been affected in some way by data deletion, such as LWL and K-NN as well as SMO. This led to better technical performance, while other techniques provided approximately better results for some classes.

CONCLUSION

Attributes are considered as the classified features of each sample data provided to data mining techniques. The strategies here have been discovered to have a significant impact on their performance after identifying the efficient attributes. For each selection method, the search algorithms and feature selection techniques have been used to find the useful features. The WBC, gender and age have been shown to have the least significant impact on the overall data. This is because of the high distance between these attributes and for the overall samples. However, datasets with relevant attributes are the only factors affecting the performance of any model. Data mining-related parameters have also been shown to have a significant effect on this objective. As a result, a variety of approaches have been implemented to concentrate on these factors and their effects. For example, for the K-NN technique with better selection and optimization, the k-value has increased the K-NN prediction from 81.4% to 84.9% with feature minimizations. It has also been concluded that a good selection of the KKT provides a 2% increase in the SMO prediction. Furthermore, it has been discovered that the SMO performs the best before and after feature selection, with roughly 84.6011% and 86.6419%, respectively. It has been demonstrated that when researchers properly choose features and method-related parameters for classification and prediction, they can easily produce improved results. The results produced have indicated that this improvement in prediction for commonly used methodologies is in addition to numerous important metrics such as MAE and RMSE. As a suggestion for future research, one can optimize this data and remove features to improve the overall classification process.

NOMENCLATURE

T	Effect parameter
$w(i)$	The weight function
i	Training dataset number
x	Testing point
w	Weight of LWL technique
Y	Class type
n	Training data numbers
f	Mapping function

Greek symbols

α	Lagrange alpha values

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

REFERENCES

[1] Pasricha SR. Anemia: a comprehensive global estimate. Blood J Am Soc Hematol 2014;123:611612. [CrossRef]

[2] Patil RR, Navghare AA. Medicinal plants for treatment of anemia: a brief review. World J Pharm Res 2019;8:701–717.

[3] Ahmad S, Banu F, Kanodia P, Bora R, Ranhotra AS. Assessment of iron deficiency anemia as a risk factor for acute lower respiratory tract infections in Nepalese children-a cross-sectional study. Ann Int Med Dent Res 2016;2:71–80. [CrossRef]

[4] Benyahmed Y, Elsanoussi KM. Effective data mining techniques performance analysis to predict anemia disease using orange tools. Fezzan Univ Sci J 2023;2:59–77.

[5] Setiawan J, Amalia D, Prasetiawan I. Data mining techniques for predictive classification of anemia disease subtypes. J Resti Rekayasa Sist Teknol Inf 2024;8:10–17. [CrossRef]

[6] Dhakal P, Khanal S, Bista R. Prediction of anemia using machine learning algorithms. AIRCC Int J Comput Sci Inf Technol 2023;15–30. [CrossRef]

[7] Kaya MO, Yildirim R, Yakar B, Alatas B. Analyzing of iron-deficiency anemia in pregnancy using rule-based intelligent classification models. Fam Pract Palliat Care 2023;8:154–164. [CrossRef]

[8] Ahmad A, Alzaidi K, Sari M, Uslu H. Prediction of anemia with a particle swarm optimization-based approach. Int J Optim Control Theor Appl 2023;13. [CrossRef]

[9] Asare JW, Appiahene P, Donkoh ET, Dimauro G. Iron deficiency anemia detection using machine learning models: a comparative study of fingernails,

palm and conjunctiva of the eye images. Eng Rep 2023;5:e12667. [CrossRef]
[10] Saputra DCE, Sunat K, Ratnaningsih T. A new artificial intelligence approach using extreme learning machine as the potentially effective model to predict and analyze the diagnosis of anemia. Healthcare 2023;11:697. [CrossRef]
[11] Zahirzada A, Zaheer N, Shahpoor MA. Machine learning algorithms to predict anemia in children under the age of five years in Afghanistan: a case of Kunduz Province. J Surv Fish Sci̇ 2023;10:752–762.
[12] El-Boghdady AM, Kishk S, Ashour MM, AbdElhalim E. Machine-learning based stacked ensemble model for accurate multi-classification of CBC anemia. Mansoura Eng J 2023;49:4. [CrossRef]
[13] Qasrawi R, Badrasawi M, Al-Halawa DA, Polo SV, Khader RA, Al-Taweel H, et al. Identification and prediction of association patterns between nutrient intake and anemia using machine learning techniques: results from a cross-sectional study with university female students from Palestine. Eur J Nutr 2024;1–15. [CrossRef]
[14] Appiahene P, Asare JW, Donkoh ET, Dimauro G, Maglietta R. Detection of iron deficiency anemia by medical images: a comparative study of machine learning algorithms. BioData Min 2023;16:2. [CrossRef]
[15] Kebede Kassaw A, Yimer A, Abey W, Molla TL, Zemariam AB. The application of machine learning approaches to determine the predictors of anemia among under-five children in Ethiopia. Sci Rep 2023;13:22919. [CrossRef]
[16] Mohammed MS, Ahmad AA, Sari M. Analysis of anemia using data mining techniques with risk factors specification. Proc Int Conf Emerg Technol INCET 2020;1–5. [CrossRef]
[17] Ahmad AA, Sari M. Anemia prediction with multiple regression support in system medicinal Internet of Things. J Med Imaging Health Inform 2020;10:261–267. [CrossRef]
[18] Yıldız TK, Yurtay N, Öneç B. Classifying anemia types using artificial learning methods. Eng Sci Technol Int J 2021;24:50–70. [CrossRef]
[19] Kou L, Sysyn M, Liu J, Fischer S, Nabochenko O, He W. Prediction system of rolling contact fatigue on crossing nose based on support vector regression. Meas 2023;210:112579. [CrossRef]
[20] Sathiyamoorthi V, Ilavarasi AK, Murugeswari K, Ahmed ST, Devi BA, Kalipindi M. A deep convolutional neural network-based computer-aided diagnosis system for the prediction of Alzheimer's disease in MRI images. Meas 2021;171:108838. [CrossRef]
[21] Beck JR, Bell JR, Hirai F, Simmons JJ, Lyon HC. Computer-based exercises in cardiac diagnosis (PlanAlyzer). Proc Annu Symp Comput Appl Med Care 1988;403.
[22] Sanap SA, Nagori M, Kshirsagar V. Classification of anemia using data mining techniques. Proc Int Conf Swarm Evol Memet Comput 2011;113–121. [CrossRef]
[23] Ahmad AA, Sari M. Parameter estimation to an anemia model using the particle swarm optimization. Sigma J Eng Nat Sci 2019;37:1335–1347.
[24] Sari M, Ahmad A, Uslu H. Medical model estimation with particle swarm optimization. Commun Fac Sci Univ Ankara Ser A1 Math Stat 2021;70:468–482. [CrossRef]
[25] Shahamat M. Synchronized components of a subshift. J Korean Math Soc 2022;59:1–12.
[26] Shouman M, Turner T, Stocker R. Applying k-nearest neighbor in diagnosing heart disease patients. Int J Inf Educ Technol 2012;2:220–223. [CrossRef]
[27] Deekshatulu BL, Chandra P. Classification of heart disease using k-nearest neighbor and genetic algorithm. Procedia Technol 2013;10:85–94. [CrossRef]
[28] Sateesh Kumar R, Sameen Fatima S. Heart disease prediction using extended KNN (E-KNN). Proc Int Conf Smart Comput Informat 2021;2:565–572. [CrossRef]
[29] Uddin S, Haque I, Lu H, Moni MA, Gide E. Comparative performance analysis of K-nearest neighbor (KNN) algorithm and its different variants for disease prediction. Sci Rep 2022;12:6256. [CrossRef]
[30] Dong X, Chen J, Zhang K, Qian H. Privacy- preserving locally weighted linear regression over encrypted millions of data. IEEE Access 2019;8:2247–2257. [CrossRef]
[31] Adnan RM, Jaafari A, Mohanavelu A, Kisi O, Elbeltagi A. Novel ensemble forecasting of streamflow using locally weighted learning algorithm. Sustainability 2021;13:5877. [CrossRef]
[32] Schneider J, Moore AW. A locally weighted learning tutorial using Vizier 1.0. Carnegie Mellon Univ, Robot Inst 2000;149.
[33] Al-Milli N. Backpropagation neural network for prediction of heart disease. J Theor Appl Inf Technol 2013;56:131–135.
[34] Kumaravel A, Pradeepa R. Layered approach for predicting protein subcellular localization in yeast microarray data. Indian J Sci Technol 2013;4567–4571.
[35] Manimurugan S, Almutairi S, Aborokbah MM, Narmatha C, Ganesan S, AlzahebHani RA, et al. An approach of CA with M-RIPPER for heart disease prediction. Research Square 2022:1–11. [CrossRef]
[36] Singh N, Firozpur P, Jindal S. Heart disease prediction system using hybrid technique of data mining algorithms. Int J Adv Res Ideas Innov Technol 2018;4:982–987.
[37] Candel D, Ñanculef R, Concha C, Allende H. A sequential minimal optimization algorithm for the all-distances support vector machine. In Progress in Pattern Recognition, Image Analysis, Computer

Vision, and Applications: 15th Iberoamerican Congress on Pattern Recognition, CIARP 2010, Sao Paulo, Brazil, November 8-11, 2010. Proceedings 15 (pp. 484–491). Springer Berlin Heidelberg. [CrossRef]

[38] Sun Z, Ampornpunt N, Varma M, Vishwanathan S. Multiple kernel learning and the SMO algorithm. Adv Neural Inf Process Syst 2010;23.

[39] Karegowda AG, Manjunath AS, Jayaram MA. Comparative study of attribute selection using gain ratio and correlation-based feature selection. Int J Inf Technol Knowl Manag 2010;2:271–277.

[40] Gagliardi F. Instance-based classifiers applied to medical databases: Diagnosis and knowledge extraction. Artif Intell Med 2011;52:123–139. [CrossRef]

[41] Schober P, Vetter TR. Logistic regression in medical research. Anesth Analg 2021;132:365–366. [CrossRef]

[42] Tiwari R, Singh MP. Correlation-based attribute selection using genetic algorithm. Int J Comput Appl 2010;4:28–34. [CrossRef]

[43] Saleem M, Ahsan M, Aslam M, Majeed A. Comparative evaluation and correlation estimates for grain yield and quality attributes in maize. Pak J Bot 2008;40:2361–2367.

[44] Huang X, Zhan J, Ding W, Pedrycz W. An error correction prediction model based on three-way decision and ensemble learning. Int J Approx Reason 2022;146:21–46. [CrossRef]