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

Research Article

Data mining technique's parameters definition and its prediction effect's based on iron deficiency dataset

Sahar J. MOHAMMED[1], Ahmed Kh. ABBAS[1], Arshed A. AHMAD[1], Mohammed S. MOHAMMED[1], Murat SARI[2], Hande USLU[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, Ripper; Attribute Selection; Prediction System; Sequential Minimal Optimization

ABSTRACT

Training dataset is not the only element affects the overall prediction system, data mining parameters have also multiple impact on their application processes. In this paper, several data mining techniques have been studied with their main parameters for effectiveness on the Anemia prediction system. For the method K-Nearest Neighbor (K-NN) process, k-value has to be defined to specify number of points to measure the distance to several types of classes. Also, Locally Weighted Learning (LWL) has a kernel value (T) which define the width of searching operation to calculate the weight function of LWL. While Sequential Minimal Optimization (SMO) has n tuples alpha values depend on the training data to satisfy the Kraush Kuhh Tucker (KKT) Condition for speeding up the prediction process. These data mining methods provided a prediction with a high performance when a better selection is optimized for each technique. Meanwhile, dataset size and attributes number have seen to have an impact on these methods performances. In this study, mining methods with feature selection methods compared in terms of proper selection of parameters and depending dataset information. Anemia system has been predicted accurately than the classical version of each method. Features for applied dataset are reduced from 11 to 8 attributes. In addition to these feature reduction and a good method's parameter selections, K-NN for example has about 3.8% increment in its prediction performances based on the proposed model.

Cite this article as: Mohammed SJ, Abbas AK, Ahmad AA, Mohammed MS, Sarı M, Uslu H. Data mining technique's parameters definition and its prediction effect's based on iron deficiency dataset. Sigma J Eng Nat Sci 2025;43(2):505−515.

INTRODUCTION

Some diseases like Anemia are considered in some research and studies as the most well-known blood illness on the planet [1]. Based on the World Health Organization (WHO), Anemia is sort of a special condition where the quantity of red platelets and, thus, the oxygen-conveying limit is insufficient to be adequate and suitable for the

*Corresponding author.
*E-mail address: dr.mohammed.sami@uodiyala.edu.iq

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

Published by Yıldız Technical University Press, İstanbul, Turkey

body's necessities [2]. Ordinary hemoglobin from one side and the other side as well as hematocrit esteems differ as indicated by age and sex. On the off chance there are beneath the limit of ordinary qualities for the age value and gender condition, then, at that point, frailty is available. The review was directed by authors to analyze 189 nations, both genders and 20 distinctive age bunches utilizing information and assets from an available dataset in 2010 that was uploaded by WHO to study on the worldwide weight of illness. They determined the worldwide iron deficiency pervasiveness as 32.9%. Iron deficiency is most regularly found in youngsters under five years of age and in ladies. The most experienced sort of weakness is iron-inadequacy paleness [3]. Since paleness, which influences personal satisfaction essentially, is both an illness and a side effect that goes with numerous genuine infections, its treatment can be basic by and large, making a right conclusion the initial move toward treatment. In practice and based on statistical evaluations, about 614 million women plus 280 million children are affected with anemia disease as in [4]. This study showed high performances when using AdaBoost regarding time, additional evaluation process, and performances by using Orange tools. This study emphasizes the potential of using machine learning to improve healthcare decisions for women and children. Several types of machine learning as also been utilized for this purpose in [5], where researchers depend on demographic data, with a focus on non-Hispanic Black females and citizens by naturalization. In this study, DT showed the best performances and could be applied as authors claimed to aid in healthcare by providing cost-effective and quality care. While other researchers and articles focused on children only and under 5 years as in [6]. In this study, authors depend on blood tests only among other anemia-related factors. For this goal, authors apply their algorithm to the noisy dataset by achieving high accuracy based on a Random Forest Tree. These performances have been increased by about 0.2% for a balanced dataset and the same version of RFT. Authors and healthcare institutions focused in general on women, especially pregnant women as explained with rule rule-based model in [7]. Authors specify the major features that affected prediction performances such as age, nutritional education status, and diversity. The authors used OneR for this objective, which obtained about a 12.4% increment in prediction accuracy for pregnant women. Also, novel hybrid approaches are provided and introduced by authors as in [8] to predict this disease regarding the related biological data. In this paper, the authors introduced an enhancement for this purpose by using Binary PSO as a feature selection. PSO was applied as a combination process with the SVM technique to offer a reliable treatment decision. The number of patients to non-patients is also considered as a main factor for some works as explained in [9]. Authors suggested 424 to 286 for the applied dataset to achieve 60% of overall anemia patients. Authors for this goal used anemia-related medical images of about 2635 samples. Image augmentation techniques were used and mixed with the mapping of Color space for data analysis using ANN. In the same field of prediction process applying an Extreme type of ML to enhance some techniques such as the Feedforward network version with a single layer in the hidden part was instructed in [10]. This model detected beta thalassemia trait and iron deficiency. Other studies deal with different age ranges as discussed in [11], when authors selected 21000 patients for 6 to 59 months with about 50% of them having anemia disease. This study was done in North-East India by using correlation-based as a feature selection among 15 attributes. Five ML techniques were applied for this purpose such as Naïve Bayes (NB). However, Random Forest among other utilized techniques achieved the highest accuracy with 15 features only. Different types of datasets with about 9747 patients were introduced as well in [12] for several anemia types. NB and other techniques are included to detect these types to provide minimum cost, low time with high performances. The variety and availability of the Anemia dataset make authors search for other related patients such as young female students as studied in [13]. Authors in this study found about 35% achieved these demonstrations by DT to identify the associations between nutrient intake (like Vitamin E and Vitamin EA) and anemia. Another study of anemia prediction based on palm-to-spot differences was conducted in [14], where authors tried also five techniques. In this study as well, NB provided higher performance for the different collected datasets from Ghana hospitals. These studies have been suggested in several countries especially as mentioned before in India, Ghana, Egypt, and also in Ethiopia as introduced in [15]. Authors in this study claimed that 40% of children are anemic especially under five years old. A feature selection was applied in this paper with RFT to provide accurate results and to specify the relationship between variables. Also, authors specify factors such as drug availability and legal restrictions that impact treatment decisions. In the medical field, computer-assisted decision-making and analysis is a common practice. A technique is developed in this study to aid and assist the specialist in the detection of several kinds of anemia. In the same context, the prior studies are reviewed and analyzed on the prediction of desired classes (types) which was undertaken [16, 17], as well as a comparison of studies that used similar methodology but used different data. Hybrid models were used in these studies [18-20]. Authors in some research devised a computer-aided system idea to present a research study in medical education, which was one of the first investigations of a programmed system that was controlled by a computer and applied for anemia diagnosis. They released the PlanAlyzer that authors applied for a diagnosis process like heart disease in 1988 [21]. This technique was designed to clarify and criticize students' approaches to identifying a common medical condition. Also, authors reported in a 1993 study that following testing and evaluation, the curriculum was studied and presented to teach the diagnosis of such a disorder as anemia and chest discomfort in the Dartmouth School of Medicine's cardiology and hematology departments. Others created computer software that provided medical advisors with a diagnostic analysis that was applied separately to 40 hematological illnesses and was designed in the Bayesian method. Different authors used the WEKA data mining tool to design a classification system to define the types of Anemia based on plenty of parameters that were taken from samples and applied several datamining techniques on it. Authors apply these techniques for 10 attributes with four types of anemia: normocytic, macrocytic, and microcytic, Microcytic (refer to thalassemia and iron deficit), macrocytic (which refers to folate insufficiency), and microcytic (iron deficiency) (renal anemia). These four anemia types are considered feature number 11 for the utilized dataset. The C4.5 decision tree method had a success rate of 99.42 percent, which was higher than the 88.13 percent success rate of support vector machines [22]. This paper is coordinated according to some steps which are explained in order: the first and second sections move towards specifying an outline of the issue that's related to paleness and a survey based on the connected writing. In the third section, the authors depict the data related to Anemia which is downloaded for this study as well as characterize pallor and lay out the strategies utilized in its findings. The fourth step sums up and talks about the outcomes that were gained in this study. At long last, section five, describes the overall work inspiration, in addition, the conceivable future review subjects are introduced.

The efficiency of several approaches especially foe selected datamining approaches on the Anemia prediction system has been examined in this article along with an analysis of their primary parameters. The K-value, which specifies the number of points to measure the distance to various sorts of classes, must be defined for the K-NN technique. Additionally, a kernel value (T) for LWL defines the search width used to determine the LWL weight function. While the KKT Condition is satisfied by SVM, which contains n tuples alpha values that depend on the training data to speed up the prediction process. This study compares feature selection and mining techniques with regard to appropriate parameter selection and dataset information. More correctly than the traditional iteration of every technique, the anemia system has been forecasted. There are now just 8 properties instead of 11 in the applied dataset's features. Apart from the feature reduction and appropriate parameter selection, K-NN, for instance, exhibits a 3.8% improvement in its prediction performances according to the suggested model.

DATASET

Samples were being taken for 539 patients with 11 features. In this case, samples have been taken from people and the blood variables 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]. A brief description of the related blood variables is provided below. The Hemoglobin (HB) located within the RBCs is a transportable protein which is composed of iron atoms. The RBC 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. WBC are responsible for safeguarding the body from infectious illnesses. HCT is a measure of the proportion of red blood cells in a given volume of blood. MCHC is the amount of the concentration of blood in a given 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 taken into account. 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 also vary according to age. For the data, it is considered that blood diseases are (1) iron-deficiency anemia, (2) deficiency vitamin B12, (3) thalassemia, (4) sickle cell, (5) and spherocytosis.

Classifier

Classification is the procedure of forecasting what class a given set of data points belongs to. Make an approximation of a mapping function (f) which related to the input data (X) into desired discrete values (y) consider as a predictive system need to be designed for this work. Data science utilizes classifiers, which are a kind of machine learning algorithm, to designate a class identification to a data input. Classifier algorithms use advanced mathematical and statistical techniques to calculate the probability of a data input being categorized in a certain manner. In this study, the following classifiers have been utilized. For defining parameters which is 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 is depend by many classifiers such as two applied classifiers in this paper.

K-nearest Neighbor (K-NN)

Simple algorithm to seek a new point which belongs to different classes based on an equal measurement. The term of neighbors is mean that any points which consider a new point should belong to an old neighbors point by adding it to them (the overall old samples). Suppose having more than one points, then by adding them, they will be classified according to the nearest class. According to Figure 1, K-NN shows that prediction altered with k-value. For example, by having k=10 which mean that 10 points are in the circle of classification and have 4 classes. K-NN will denote the new points to any near and counted class. As shown in Figure 1 each class has new added points to the samples (2, 4, 3 and 1 point) belongs to (Disease type 1, 2, 3 and type 4) respectively. Starting with k-value = 10, means selecting 10 points near to a center point, according to Figure 1, the majority points belong to the disease type 2 (4 points), therefore; K-NN Can be written as a class 2 for k=10. When k=13, a new point added to the old samples to be 19 points. But in this k-value, K-NN is shown to be a disease type 4 because of the new added points. In this paper, selecting k value was important to give the optimal solutions by using the cross validation Technique. To find the minimum distant between points, it's done by applies two types of distance calculating methods (Euclidean and Manhattan). In similar types of measurement factors, K-NN can be applied for disease prediction such as heart disease in [26] or combining this algorithm with a genetic algorithm as in [27]. Several articles proved that K-NN when compared with another type of machine learning techniques have been providing a better result such in [27-29].

Locally Weighted Learning (LWL)

When dataset collected is unknown in shape or patterns, the closet points, that wanted to be tested can be estimated according to nearest samples. By weighting the nearest points to the predicted one, it should assist the whole process. LWL is not a parametric method, its depends on the training data. In Figure 2 the simple explanation of LWL showed how is the measuring process between tested patient dada parameter 1 with the disease type of Anemia.

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

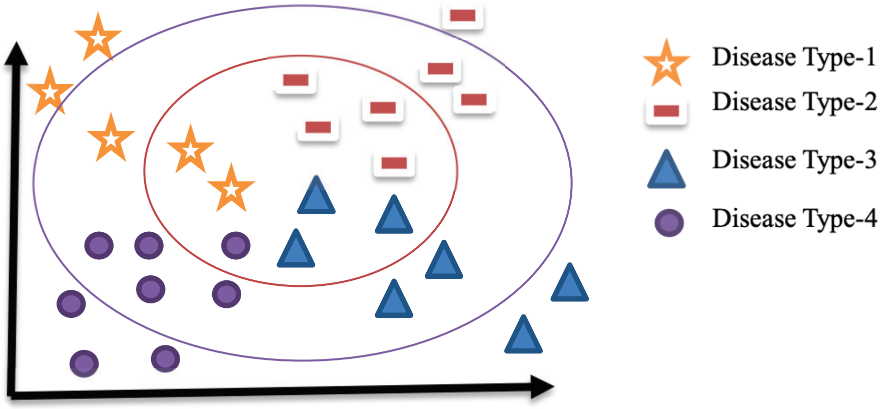


Figure 1. K-NN algorithm explanation according to k- value.

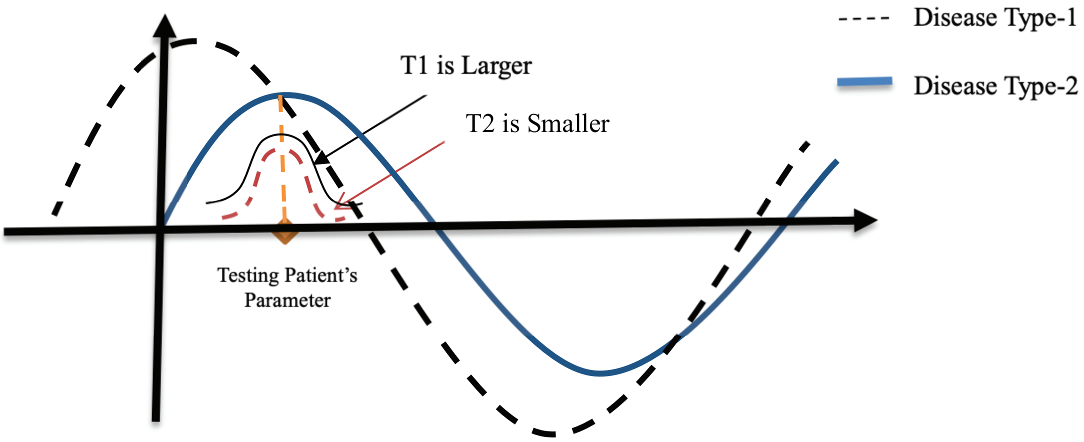


Figure 2. LWL algorithm explanation according to T value.

$$\text{Kernel Weight (i)} = \exp\left\{\frac{x^{(i)}-x}{2T^2}\right\} \quad (1)$$

Where i is the number of training date and x is the testing point, which showed the affected of T value on weight and distance calculating process, as shown in Figure 2, when $T=T_1$, LWL has more points near to the tested one. While T_2 Provided less 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 hyper parameter. The denominator of kernel weight defines the distance state between training and testing data, if distance is close to the training date, then denominator = zero, while the weight kernel in general equal to 1 or near to the highest value of kernel. If the distance is big the denominator is bigger, then the w(i) is near to zero. The differences between standard and LWL regression mode is shown in Equation 2 to find the optimal value as in [30-32]. Where x is the training data, w is the weight and Y is the class type.

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

Ripper

It refers to minimize, error by incremental repeated methods. Ripper modifies the performances of Decision Tree (DT) by evolving multiple iteration which can be presented by three steps: (Growing, Pruning and Optimization). The first step, applying the same steps of the DT to make the corrected path according to data and its attributes. All attributes added and checked by Growing DT process until no longer 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 showed the steps of modifying C5.0 steps by using this model.

Sequential Minimal Optimization (SMO)

It is a sequential process to optimize the smallest sub functions for each iteration steps. In this technique, alphas value has been supplied from SMO to satisfy the constraint of the required problem. SMO alphas are denoted as Lagrange multipliers which can be calculated easily. These multipliers should be identified firstly before working or applying any SMO Steps. And depending on training data numbers, n values of SMO Lagrange will be selected from all these defined alphas. which is considered as the smallest sub-problem for the given main or required problem. For any training dataset (n-values), n(n-1) possibilities are

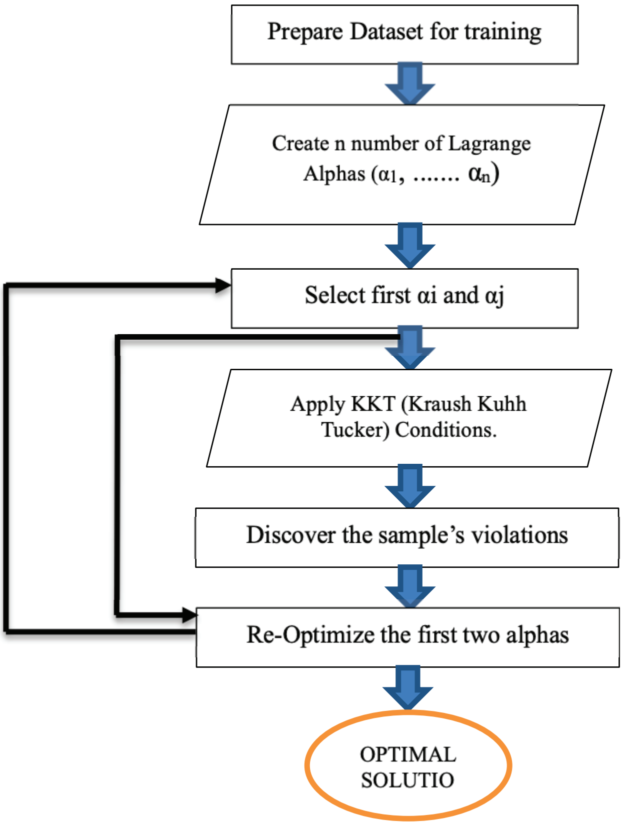

Prepare Dataset for training
Create n number of Lagrange Alphas (α_1, ……. α_n)
Select first αi and αj
Apply KKT (Kraush Kuhh Tucker) Conditions.
Discover the sample's violations
Re-Optimize the first two alphas
OPTIMAL SOLUTIO

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

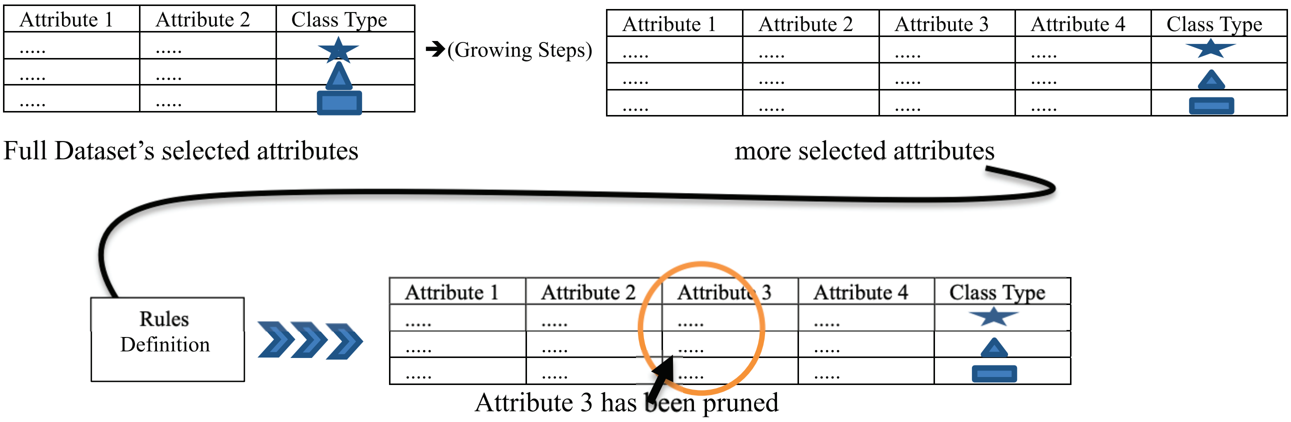


Figure 3. RIPPER steps to gain optimal solutions with several epoch numbers.

derived, but at each iteration two possible values have been selected to accelerate the convergence. By applying KKT condition to discover the violations samples which need to be optimization in the next steps (next iterations). after finding a violations of training samples according to αi and αj, these two alphas will be re-optimized while keeping the rest of alphas constant. Repeating the previous steps until all alphas satisfies the KKT Conditions and no other violation of samples are found. Any larger problem is being sub divided into a small sub problem with two alphas to speed up the calculation process as mentioned in [37-39]. Figure 4 shows the anemia prediction steps based on Anemia dataset and considering the back version of SVM (SMO).

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 moving as steps. It differs from DT and K-NN in memorizing all samples instead of building any model that is stored in memory and comparing new samples with saved samples. Providing decision was based on K most like samples not like distance metrics 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 is also having 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, sensitive to high dimensional features, and sensitive to imbalanced classes when one class type is more repeated than another as mentioned in [40].

Logistic Regression Model (LLRM)

This approach is widely used for binary classification tasks, it uses a function known as logistic to model the relation and probability for a given data belonging to a certain class. 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 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 related with features to express the moving process of prediction, larger features provide harder prediction system. Minimization of these attributes are important to select the most impact features on prediction system in addition, to provide time prediction minimization as mentioned in [38]. The high measurement information makes testing and preparing of common classification strategies difficult. Concurring to property determination handle which is done by a few strategies like: - Attributes Evaluation, Correlation, Pick up Ratio and Principal Component. Which was also done by applying the reasonable looking strategies related to each of the choice process such as: - Best First and Ranker. From all of the over selecting strategies, WBC, Gender and Age were the least affect features on the overall data. Even feature selection techniques have been differences such as explained in [39], which provided the impact features of a diabetic India database. In some articles, further selection techniques have been applied by mixing Genetic Algorithm (GA) with it as in [42] or even apply it for different fields not just mining algorithm such in agriculture [43] and biomedical [44].

RESULTS AND DISCUSSION

Samples were being taken for 539 patients with 11 features for the first time of classification and before features minimization. Applied techniques as explained before are the Instance Based Classifier (K*), LWL, RIPPER, SMO, K-NN classifier and Logistic Regression Model (LLRM). Each of applied techniques are 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 of these techniques under the same conditions for 11 attributes and 10-fold validation before feature selecting utilizing. Table 1 shows the overall accuracy according to 11 attributes for above techniques. This table shows that SMO has the greatest accuracy among all of applied techniques. This is due to kernel transformation which assist this technique to linearize data. While LWL was the worst classification technique due to the weighted measurement of each sample according to the neighbor (required) sample which is sort of approximately value. For this reason, data should be first welly identified and minimized to be more accurate for these types of techniques to provide more accuracy to Anemia data classification.

In addition, the same table shows the same 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 features

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

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

minimization to make data more linear before dealing with these types of techniques. According to SMO, weighted samples would be more accurate with a well separated data so distance will be truly classified. MAE refers to the error between each paired samples in the same data, which provided that these data should be carefully minimized or edited to be welly classified with a minimum absolute error. While RMSE pointed to the differences between the desired prediction samples and truly corrected samples. Which is also lead to that some of data's parameters are 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 technique's performances. 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 attribute selection process which is done by some methods like: - Attributes Evaluation, Correlation, Gain Ratio and Principal Component. Which was also done by applying the suitable searching methods related to each of the selection process such as: - Best First and Ranker. From all of the above selecting methods, WBC, Gender and Age were the least impact features on the overall data. This is might 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 calculating parameters but for 8 attributes to study this impact.

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 (SMO) in data classification after omitting three number of attributes. While Figure 6 shows the overall MAE and RMSE for the used data. Which is also provided the same results.

Another study has been done for this work to prove system enhancement after selected the most significant data. Figure 7 below shows some classes' performance according to compared parameters such as TP Rate, FP Rate, Precision and PRC. Which guided this work to the main points that makes these data need to be truly defined and specified.

Figure 8 shows SMO parameters for the first type and the third type of anemia after features selecting and deleting three of the unusual attributes, and also, shows that SMO performances are changed to get a welly define samples than before after deleting some of the unrequired features. This is due to most datamining techniques are distance dependent of data.

This results were done for all types of Anemia classes and for both cases before and after feature selections. Also, this study was also done for all of the mentioned technique to give a brief knowledge about features affection on datamining techniques and if this affection was equal to all techniques or not. All of the studied techniques were sort of affected by data deleting like SMO such as LWL and K-NN. While other techniques approximately provided better results for some classes already which also led to better technique performances.

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

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

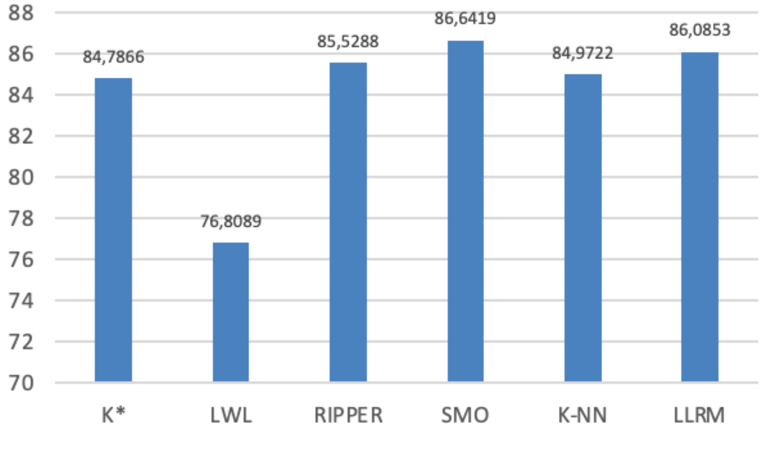


Figure 5. Overall system accuracy for the applied techniques.

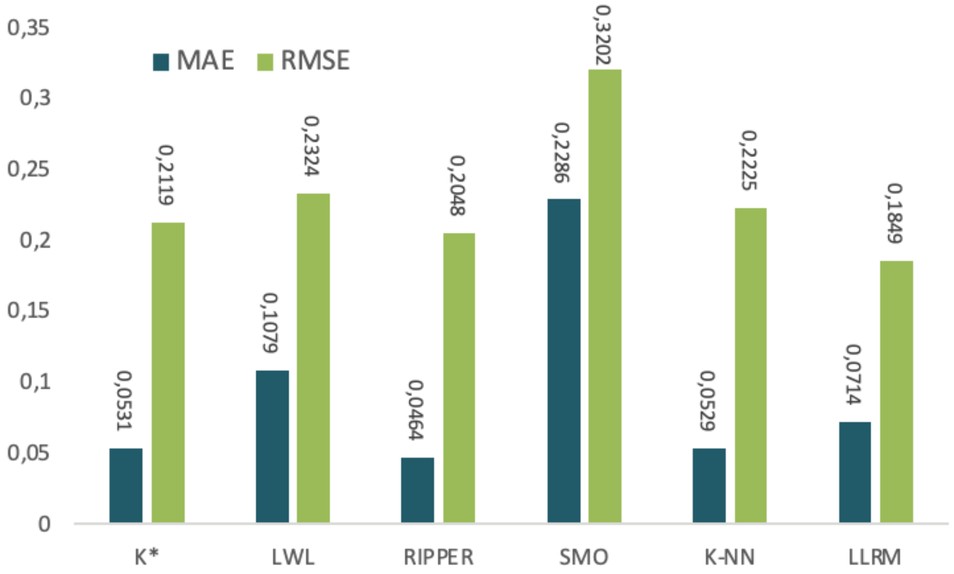
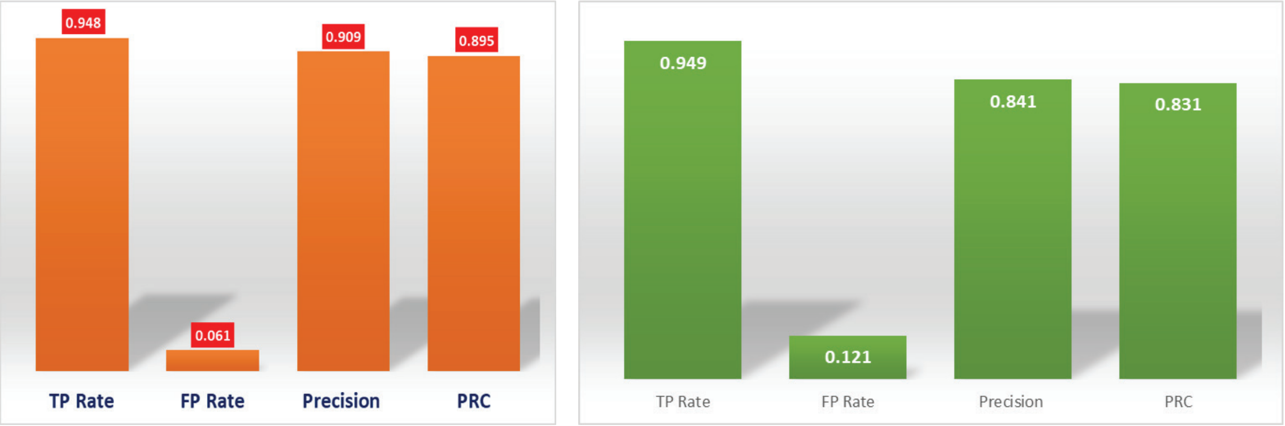
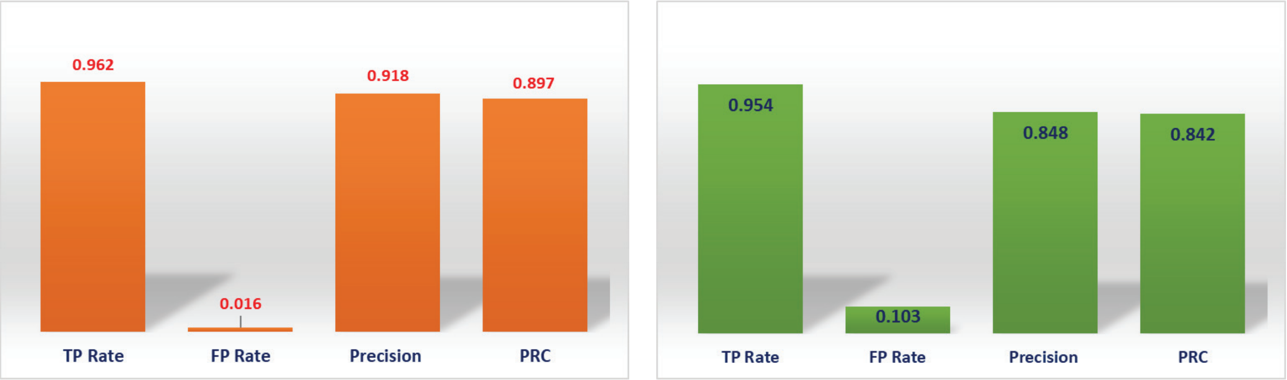


Figure 6. Overall system MAE and RMSE for the applied techniques.



(a) (b)

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



(a) **(b)**

Figure 8. SMO parameters for the (a) first type and (b) third type of anemia after features selecting.

CONCLUSION

Attributes consider as the characteristics of each sample that data provided to datamining techniques to be classified. For this reason, it had a significant impact on the techniques performances after specifying the useful attributes than others. Useful attributes were determined according to feature selection methods and searching algorithms that related to each of the selection methods. Which provided that WBC, gender and age had the least significant impact on overall data. This is due to the high distance between these attributes and for the overall samples. However, dataset with relevant attributes are not the only effective factors on any system performances. Data mining related parameters also demonstrated as a significant impact on this purpose. For this reason, multiple types of techniques are applied to focus on these parameters and its affection. For example, k value for K-NN technique with a better selection and optimization enhanced the prediction of K-NN from 81.4% to 84.9% with feature minimizations. Also, a good selection of KKT provide 2% increment in SMO prediction system. In addition, SMO had the best performances before and after feature selecting with about 84.6011% and 86.6419% respectively. Which demonstrated that researchers when carefully selected features and method's related parameters for classification and prediction purpose better performances will have obtained easily. Table 1 and 2 shows this enhancement in prediction performances for the overall utilized techniques in addition to several important metrics like MAE and RMSE. As a suggestion for a future work will be optimize these data and the deleting attributes to be benefit to the overall classification process.

T is the affect parameter to determine the width of the kernel function. Where the k(i) is the weighted function (which mean the distances between the tested point and all other points in the training date) to provide a weight as in Equation 1.

Where i is the number of training date and x is the testing point, which showed the affected of T value on weight and distance calculating process, as shown in Figure 2

weight kernel in general equal to 1

Where x is the training data, w is the weight and Y is the class type.

And depending on training data numbers, n values of SMO Lagrange will

For any training dataset (n-values), n(n-1)

NOMENCLATURE

T	Effect Parameter
$k(i)$	The weighted function
i	Training dataset number
x	Testing point
w	Weight of LWL technique
Y	Class type
n	Training data numbers

Greek symbols

α	Lagrange Alphas

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;123:611−612. [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, Murat SAR. 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, Sarı M. Parameter estimation to an anemia model using the particle swarm optimization. Sigma J Eng Nat Sci 2019;37:1335–1347.

[24] Murat SAR, Ahmad A, Hande Uslu. 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]