Kafrelsheikh Journal of Information Sciences ISSN (Online): 2535-1478, ISSN (Print): 2537-0677 Volume 4, Issue 2, 2023, PP. 1–9
Date of publication 25 November 2023

# Anemia Diagnosis And Prediction Based On Machine Learning

AbdulRahman Khawaga [1], Eman A.Shehab [2] Sara A.Shehab [3]
[1] Student, Bioinformatics DP, Faculty of Computer and Artificial Intelligence, University of Sadat City.
[2] Faculty of Computer and Artificial Intelligence, University of Sadat City
[3] Faculty of Computer and Artificial Intelligence, University of Sadat City

Article Info
**Keywords:**
*Anemia.*
*Hemoglobin*
*Classification*
*Machine Learning.*
*Random Forest.*
*Decision Tree.*
*Logistic Regression.*
*Support Vector Machine.*
*Multilayer Perceptron.*

## Abstract

*The extraordinary developments in the health sector have resulted in the substantial production of data in daily life. To get valuable information out of this data—information that can be used for analysis, forecasting, making suggestions, and making decisions—it must be processed. Accessible data is converted into useful information using data mining and machine learning approaches. The first challenge for medical practitioners in developing a preventative strategy and successful treatment plan is the timely diagnosis of diseases. Sometimes, this can result in death if accuracy is lacking. In this study, we examine supervised machine learning methods (Decision Tree, Multilayer Perceptron "MLP", K-nearest neighbors "KNN", Logistic Regression, Random Forest, and Support Vector Machine "SVC") for anemia prediction utilizing CBC (Complete Blood Count) data gathered from pathology labs. The outcomes demonstrate that the Random Forest, Multilayer Perceptron "MLP", Decision Tree, and Logistic Regression techniques outperform KNN and SVC in terms of accuracy of 99.94%.*

## 1. Introduction

Anemia is one of the most common blood disorders worldwide and can affect all types of people [1]. It is important to keep in mind that the majority of blood disorders are caused by abnormalities in particular genes and can be passed down through families. Blood disorders can also be caused by medical conditions, including drug use and lifestyle choices. According to reports, the most popular blood condition affecting humans is anemia [2]. Anemia is a condition when there are either too few red blood cells or too little hemoglobin in them. A person's blood's ability to transport oxygen to the body's tissues will be reduced if they have insufficient or abnormally few red blood cells or if they have insufficient amounts of hemoglobin, which is required to deliver oxygen. This causes symptoms like weakness, exhaustion, feeling dizzy, and shortness of breath. Age, sex, smoking habits, and the status of pregnancy all affect the ideal hemoglobin concentration needed to meet physiologic needs [3].

### 1.1. Prevalence of Anemia

According to the World Health Organization (WHO), one-quarter of the world's population suffers from anemia, which has a particularly negative effect on children between the ages of 6 and 59 months and pregnant women [4]. The number of people who suffer from anemia grew globally by 0.3% in 2019 to 29.90% [5],it affects pregnant women and children at rates of 47.4% and 41.8%, respectively [6]. As a result, societies where anemia is common suffer enormous economic losses [7]. Although the disease's numerous symptoms make it difficult for people to diagnose the disorder due to its hidden nature, it is a significant and serious problem regardless [8]. To decrease anemia's prevalence, it is essential to spread the necessary knowledge of its causes and symptoms to try to treat this disease as much as possible [9].

### 1.2. Causes and Types of Anemia

Anemia has many causes; Iron deficiency is thought to be the main cause of anemia worldwide, but other nutritional deficiencies (such as folate, vitamin B12, and vitamin A deficiency) can be the main causes. Additionally, anemia may result from acute or chronic inflammation, parasite infections, inherited or acquired disorders that disrupt the formation of red blood cells, the survival of those cells, or hemoglobin manufacturing [3]. For example, the defects in hemoglobin or the synthesis of abnormal hemoglobin cause different and more dangerous types of anemia, such as Sickle Cell Anemia and Thalassemia[10] .

### 1.3. Symptoms and Diagnosis of Anemia

The first and most crucial step in diagnosis is to identify the clinical symptoms and indicators of anemia. However, some people won't exhibit any symptoms, and in certain cases, a diagnosis will be made based on unexpected laboratory results. Patients will have symptoms related to volume loss in acute situations, including dizziness, syncope, and hypotension [10]. Another symptom that can be an indicator of anemia is pale or yellowish skin, which might be more obvious on white skin than on black or brown skin. Although chronic anemia can be asymptomatic, it can also worsen comorbid conditions like angina, heart failure, chronic kidney disease, and chronic obstructive pulmonary disease [11]. Complete blood count (CBC), ferritin, PCR (polymerase chain reaction), and hemoglobin electrophoresis are the four primary tests used to identify anemia disorders[12] [13].The most common blood test utilized to evaluate general health and identify a variety of disorders [8], such as anemia, infection, and leukemia, is the CBC test. Hemoglobin (Hb), red blood cells (RBC), hematocrit (HCT), mean corpuscular hemoglobin (MCH), mean corpuscular volume (MCV), and more are among the almost 15 parameters that a full blood count test examines [14].

This research's primary goal is to construct a model employing several machine learning algorithms and evaluate those algorithms' performances, considering assessment criteria for the prediction of anemia using a complete blood count (CBC). The main contribution in this paper is as follows: -

1. Predicting the anemia percentage helps in avoiding many diseased.
2. Many machine learning models are proposed for predicted the anemia.
3. The proposed model achieved 99.94% accuracy.

The remainder of the paper is organized as follows: The existing relevant work is outlined in Section 2. Details about the dataset will be presented in Section 3. The methodology and models used are provided in Section 4. Then there are discussions about the results in Section 5. Eventually, you can reach the conclusions of this study in Section 6.

## 2. Related Work

Machine learning (ML) techniques have been widely used to detect various diseases over the past ten years. This makes an early diagnosis easier and raises the likelihood of survival.

There are many different types of machine learning algorithms and techniques, including support vector machines (SVM), naive bayes (NB), decision trees (DT), k-nearest neighbor (KNN), multilayer perceptron (MLP), hybrid classifier machine learning, average ensemble (AE), genetic algorithm convolutional neural network (GA-CNN), genetic algorithm stacked encoder (GA-SAE), random forest (RF), and support vector machines. According to [6] [12] [15], which predicted data in the form of a complete blood count (CBC) and built a model to diagnose anemia, several papers have been published on the application of machine learning to categorize various types of anemia. The authors said in [9] that in the past, computerized systems were created with the assistance of medical professionals and afterwards converted into algorithms. This approach took a lot of time, though. Association rule mining and decision tree methodologies have produced major solutions to issues in the health industry.

By utilizing machine learning approaches (ANN, SVM, NB, RF, and KNN) to predict anemia and malaria in children, Boubaker Sue et al. [16] concentrated on investigating social aspects that are thought to be significant contributors in children's health. The data set utilized for the Demographic and Health Surveys (DHS) carried out in Senegal in 2015 and 2016 provided the data for this investigation. The ANN predicted anemia with an accuracy of 84.17% and malaria with a precision of 94.74%.

In [17], WEKA is employed to create an appropriate classifier for the creation of a mobile application that can anticipate and diagnose remarks made in hematological data. The J48 and Naive Bayes classifiers were put to the test against neural network classification techniques by the authors. The findings reveal that the J48 classifier has the highest level of accuracy.

A decision support system was developed by authors in [18] to diagnose iron-deficient anemia using a decision tree algorithm. This system takes into account ferritin, serum iron- list capacity, and the three hematological parameters. The evaluation was grounded in data from 96 cases, and the issues were favorably compared to the doctor's decision.

Estimating hemoglobin levels is a crucial stage in any blood analysis work [19], and it also establishes whether a person is anemic. In research [15], hemoglobin levels were determined, and anemia was identified using blood test features and a machine learning model. 9004 records make up the dataset, of which 6753 were utilized for training and 2251 for testing. Three different machine learning algorithms—DT, NB, and NN—as well as a hybrid classifier, which combines all three methods, were used. Additionally, the performance of the methodology was evaluated using the MAE and RMSE approaches. According to the MAE findings, the hybrid classifier had an accuracy of 0.996% and the best RMSE value of 0.015 [15].

In [20], authors used machine learning algorithms such as Random Forest, SVM and others to predict whether the patient is anemic or not. They developed a classification-based ML model and used it to forecast a patient's anemia using crucial CBC test data.

As can be seen until now, the predictive power of machine learning has increased incredibly, so many of the studies have used several ML methods such as Support Vector Machine, Random Forest, Naïve Bayes, Decision trees and Multi-Layer Perceptron to build a model to predict a person is anemic or not. So, in this paper, we are going to use some of these algorithms to make a predictive model and choose the best algorithms according to their accuracy in order to use them or combine it to predict another type of anemia or other diseases soon.

## 3. About the dataset

In this section, an introduction describes the dataset used and its characters and gives information about them. The dataset used in this paper was collected from Kaggle [21]. This anemia prediction dataset is based on CBC test data of 8544 records. It consists of 5 parameters and one label: Gender, Hemoglobin, MCH, MCHC, MCV, Result (Label) and here is a brief description of them:

**Gender**: Given that male and female blood parameters and limitations alter and change, gender is an essential parameter that must be considered [3]. In this dataset, 0 is encoded as male and 1 as female.

**MCH**: The abbreviation "MCH" stands for "mean corpuscular hemoglobin." It is the average concentration of hemoglobin, a protein that transports oxygen throughout your body [3], in each of your red blood cells.

**MCHC**: MCHC, which stands for "mean corpuscular hemoglobin concentration," is a metric comparable to MCH.

The average quantity of hemoglobin in a collection of red blood cells is measured by MCHC. Both MCHC and MCH may be used by a physician to identify anemia [20].

**MCV:** Mean corpuscular volume, or MCV, is a medical term. In essence, this blood test calculates the red blood cells' average size. We may learn whether our red blood cells are too small or too big with this test, which can indicate any blood illness like anemia.

**Hemoglobin:** This measurement reveals how much oxygen is in our blood. It is essentially a protein that can transport oxygen throughout the body. It is also a crucial factor in the prediction of anemia. Anemia is commonly described as having hemoglobin levels of less than 13.5 g/dl in males and less than 12.0 g/dl in females [20].

**Result**: Result here means whether the person is anemic or not (0 for not anemic and 1 for anemic).

All parameters in our dataset are numerical. Table 1 shows a random sample of the dataset.

Table 1 Random dataset samples

| Gender | Hemoglobin | MCH | MCH C | MCV | Result |
|---|---|---|---|---|---|
| 1 | 13.8 | 33 | 96 | 31.7 | 0 |
| 0 | 12.6 | 34.4 | 72.8 | 25.1 | 1 |
| 0 | 11 | 34.9 | 83.3 | 29.1 | 1 |
| 0 | 13.4 | 33.5 | 79.1 | 26.5 | 0 |
| 1 | 15.4 | 33.8 | 94.1 | 31.7 | 0 |
| 0 | 10 | 33.3 | 93 | 31 | 1 |
| 0 | 12.8 | 33.6 | 87.2 | 29.2 | 1 |

# 4. Proposed Methodology

The proposed methodology is divided into three phases: I) Data pre-processing. II) Classifying using different ML algorithms. III) Measure performance. This section is a summary of these phases. See figure 1.

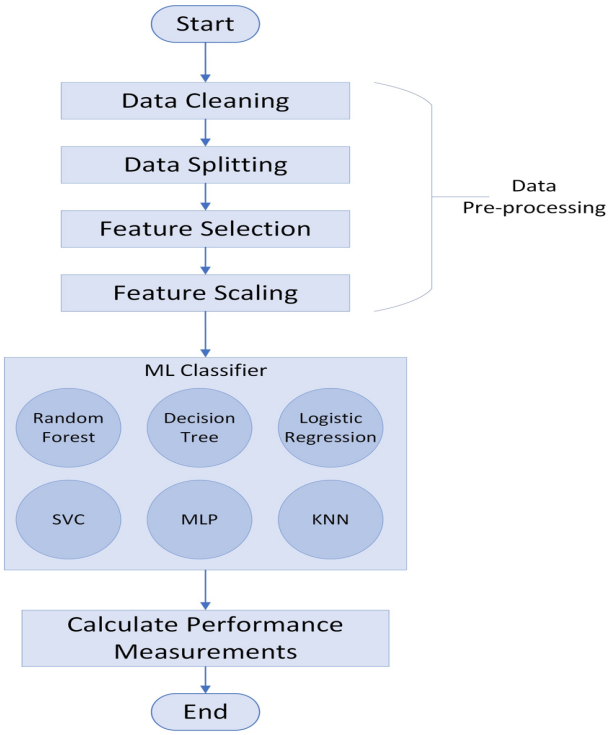


Fig. 1.Proposed Methodology

## 4.1 Data Preprocessing

The data in its normal condition is an unbalanced dataset as the majority of most medical datasets [22],and as a result, it will affect the prediction models will be used as the models that will be biased towards the majority class only, which here is not anemic (0), and as a result, the predictive model built won't give a good accuracy in the prediction of the minority class, so with the use of random under sampling we have made a balanced dataset with 5808 records with 50% for anemic results and 50% for not anemic results. In addition, missing values were added to the dataset to help the model train missing values to get a high precision and accuracy.

## 4.2 Data Cleaning

The duplicate data and the missing data are examined at this stage of the pre-processing procedures.Fig.2 below shows the sum of missing data in the dataset for each column. As shown in the figure, there are 35 missing values.

```
Gender        0
Hemoglobin    8
MCH          11
MCHC          7
MCV           9
Result        0
dtype: int64
```

Fig.2. Number of missing data in each column in the dataset before cleaning

There are numerous approaches for handling missing values, such as by substituting the value with the attribute's median, mean, or mode. In this study, the SimpleImputer [23] was used with the mean strategy, as the mean value for each feature column is used to fill in for a feature's missing value. Additionally, it was made sure that no values were repeated or missing, as shown below in Figure 3.

```
Gender          0
Hemoglobin      0
MCH             0
MCHC            0
MCV             0
Result          0
dtype: int64
```

Fig.3.Number of missing values after cleaning

Now that the data is cleaned, we can progress with our study.

## 4.3 Data Splitting

When a machine learning model fits its training data too well and cannot reliably fit fresh data, it is said to be overfit. In order to avoid overfitting, data splitting is widely utilized in machine learning. A machine learning model often divides the initial data into two or three. The three sets that are usually utilized are the training set, the validation set, and the testing set. The piece of data used to train the model is known as the training set. In order to improve any of its parameters, the model must observe and learn from the training set. The validation set is a data collection of examples used to alter the settings for the learning process. The objective of this data set is to rate the model's accuracy, which can aid in model selection. The data set that is tested in the final model and contrasted with the earlier data sets is known as the testing set. The testing set serves as an assessment of the chosen algorithm and mode.
Firstly, we must divide the dataset into two parts to make it easier to deal with the data. One part is called X which consists of the 5 parameters of the CBC data (Gender, Hemoglobin, MCH, MCHC, MCV) and the other part is called y and it's our label (Result). Then we can split our data. In this study, the data were split into two parts: train and test with a ratio 70:30, as shown in figure.4.

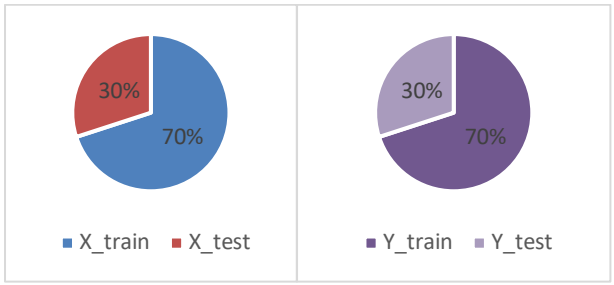


Fig.4.Data Splitting

## 4.4 Feature Selection

Finding the optimum collection of features that can be utilized to build practical models is the goal of employing feature selection strategies in machine learning. It includes determining each input variable's link to the target variable using a set of evaluation criteria before choosing the input variables with the strongest relationships. Feature selection helps to reduce noise in the data and is used to decrease the size of the dataset, increase decision accuracy, and shorten the time needed to complete the ML training process. The filter approach, wrapper approach, embedding approach, and hybrid approach are the four basic feature selection methodologies [24]. One of the most common machine learning techniques is random forests. They are highly effective because they often have strong predictive accuracy, little overfitting, and are simple to understand. The fact that it is simple to determine the significance of each variable on the tree decision contributes to this interpretability. In other words, it is simple to calculate the percentage of the choice that each variable contributes.
Using a random forest to choose features falls under the heading of embedded approaches which combine between wrapper and filter methods.

So, in this study, Random Forest (RF) is used as a feature selection method. As a result, Figure 5 shows the importance of each feature in the dataset in a visualization chart as shown below, the most important feature is the one with index 1, which is hemoglobin and comes after it the gender with index 0 with the rest of the features having very low importance to the model. So as a result of the feature selection step, hemoglobin and gender were chosen to work with their data in the prediction model.

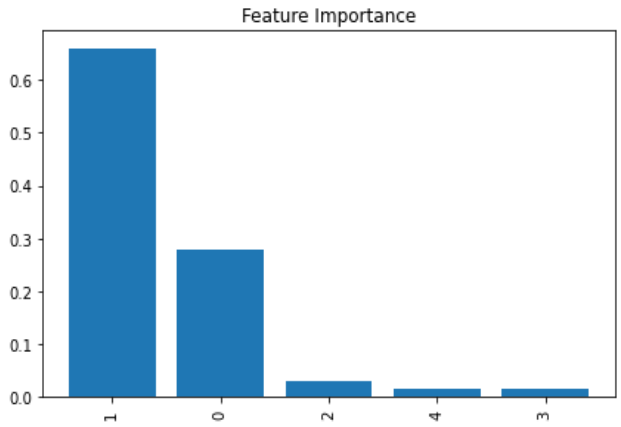


Fig.5. Feature Selection by using Random Forest (RF)

## 4.5 Feature Scaling

A Feature scaling step is a data preprocessing technique used to normalize the range of features or variables in a dataset to a similar scale. This prevents features with higher values from dominating the model and ensures that all features contribute equally to it. When working with datasets where the features have multiple ranges, units of measurement, or orders of magnitude, feature scaling is crucial. Standardization, normalization, and min-max scaling are common methods for feature scaling. The data may be shifted to a more uniform scale by using feature scaling, making it simpler to create precise and efficient machine learning models.
A Standard Scaler (Standardization) is used in this study, which uses the following equation:

$$Z = \frac{x-\mu}{\sigma} \qquad (1)$$

Where z is the new data, x is the old data, $\boldsymbol{\mu}$ is the mean of the feature values, and $\boldsymbol{\sigma}$ is the standard deviation of the feature values.
Now that the data has been cleaned, we can work with ML algorithms and start working with the models,and find predictions.

## 4.6 Machine Learning (ML) classification algorithms

Machine Learning (ML) is concerned with creating algorithms that let computers learn on their own without the need to program each rule to make a choice or extract a certain pattern. In supervised learning, algorithms build a model using inputs (features) and outputs (targets) with the goal of producing precise predictions for future data inputs. Algorithms learn correlations and patterns that they utilize to anticipate outcomes from fresh data [25]. Classification is one of the most significant and popular supervised ML approaches. Binary and multiple classification are the two forms that depend on the total number of outputs [26]. Binary classification was implemented in this study because we can diagnose and predict whether a person will be anemic or not. Six classifiers were used in this study:

## 4.7 Random Forest (RF)

Random forest (RF) is a classification method based on "growing" a group of classifiers with a tree-structure. Each classification tree in the forest is used to categorize a new individual using the characteristics of that person. Each growing tree provides a categorization (or "voting") for a class label, and the trees are constructed at random. The choice is based on the votes cast by the majority of the forest's trees [27].

A random forest classifier with a number of estimators of 1000 was used in this study.

## 4.8 Decision Tree

Decision tree is one of the supervised learning approaches that is used for classification and regression. It is organized hierarchically, with a root node, branches, internal nodes, and leaf nodes. Each branch node denotes a decision, while each leaf node denotes a choice among several options. The goal of data mining is to uncover as much hidden information as possible; hence, this approach is extensively researched. The subject of medicine is one that still has a lot of unexplored territory [28].

## 3.9 Support Vector Machine (SVM)

Support vector machines (SVM) are a frequently used kernel-based supervised machine learning approach for classification issues. It works by maximizing the margin between the classes, the SVM algorithm builds a hyperplane that properly divides the training observations according to their class labels. Depending on which side of the hyperplane a test observation is on, SVM allocates it to a class [29]. Among other algorithms, its contribution is related to medicine and medical personnel, improved illness diagnosis, better treatment recommendations, and medication dosages.

SVM can also work as a regressor, but in this study we are classifying, so it will not be used, so our concentration will be only on Support Vector Classifier (SVC) with a 'rbf' kernel, 1500 maximum number of iterations, 'auto' gamma, and C=1.0.

## 4.10 Logistic Regression (LR):

LR is a supervised classification technique that fits data to a logistic curve to estimate the probability of an occurrence. The result is evaluated using a binary variable. In logistic regression, several anticipated variables—some of which may be categorical or numerical—are employed. LR is often employed in the social sciences and in healthcare. It is also often used in marketing to determine whether people are likely to purchase a product [30].

A logistic regression model with a maximum number of iterations of 10000 was used in this study.

## 4.11 K-nearest neighbors (KNN):

K-NN: The k-nearest neighbors (k-NN) classifier belongs to the supervised learning family of algorithms and is a reliable and adaptable classifier. Because it does not make any explicit assumptions about the distribution of the data, k-NN is a non-parametric method. This method classifies new instances using a similarity metric and saves all of the existing examples. A case is categorized by a majority vote of its neighbors, and the case is then put into the class that is most prevalent among its k closest neighbors as determined by a distance function.

A KNN classifier is used with a number of neighbors of 5 in this study.

## 4.12 Multilayer Perceptron MLP:

The MLP classifier is frequently referred to as a multilayer representation perceptron classifier, which denotes a neural network itself. MLP accomplishes the classification job based on the underlying neural network, in contrast to other classification algorithms like SVM and NB. Three layers of nodes make up the multilayer sensory structure of MLP: an input layer, a hidden layer, and an output layer. The input layer is the top layer. The first layer receives the training data and multiplies it by weights that have been initialized at random. After that, biases are introduced, and the final product is then given an activation function. Each layer's input data comes from the layer before it, with the exception of the first layer, and the process is repeated with the output being transferred to the following layer [31] [32].

Performance Measurements:

After finishing classifying with all the proposed algorithms, the output of each classification algorithm is visualized and summarized in a confusion matrix. A confusion matrix is a table used to describe how well a classification system performs.

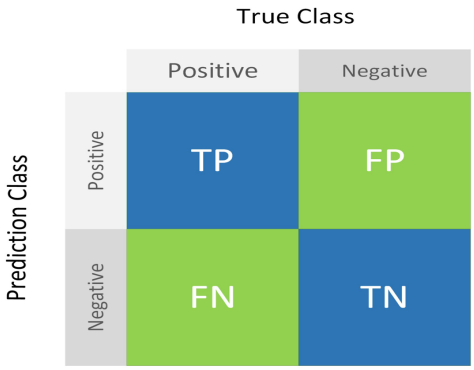


Fig.6. Confusion Matrix

As shown in Figure 6, four fundamental properties make up the confusion matrix, which is used to provide the classifier's measurement parameters. These are the four parameters:

TP (true positive): Total number of cases that were predicted yes and were correctly predicted (in our study, means that they really have the disease and the model predict that correctly).

TN (true negative): Total number of cases that were predicted no and were correctly predicted (in our study, it

means that they don't have the disease and the model predicted that correctly).
FP (false positive): Total number of cases that were predicted yes and were a wrong prediction (in our study, it means that they don't really have the disease, but the model predicted that they did).
FN (false negative): Total number of cases that were predicted no and was a wrong prediction (in our study, it means that they really have the disease, but the model predicted that they don't have it).
Classification accuracy, precision, recall, specificity, and F1-score are called performance metrics. It is calculated for comparison and assessment of our suggested methods and is calculated as shown below in equations (2), (3), (4), (5), and (6), respectively.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (2)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (3)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (4)$$

$$Specifity = \frac{TN}{(TN+FP)} \quad (5)$$

$$F1 = \frac{2\,.\,precsion\,.\,recall}{precision+recall} \quad (6)$$

In addition, the receiver operating characteristic curve (ROC) curve was also plotted, and the area under the curve (AUC) was calculated.

## 5 Results and Discussions

The proposed work was implemented using Python. After implementing all the previous steps in our study, we now present all the results and performance metrics of each ML classifier used. The confusion matrix of the proposed algorithm is shown in Figure 7. And the ROC is shown in Figure 8. To indicate the improvement of the proposed work the results are compares to previous work in table 2. The comparison proved the ability of proposed work to predict the anemia effectively.

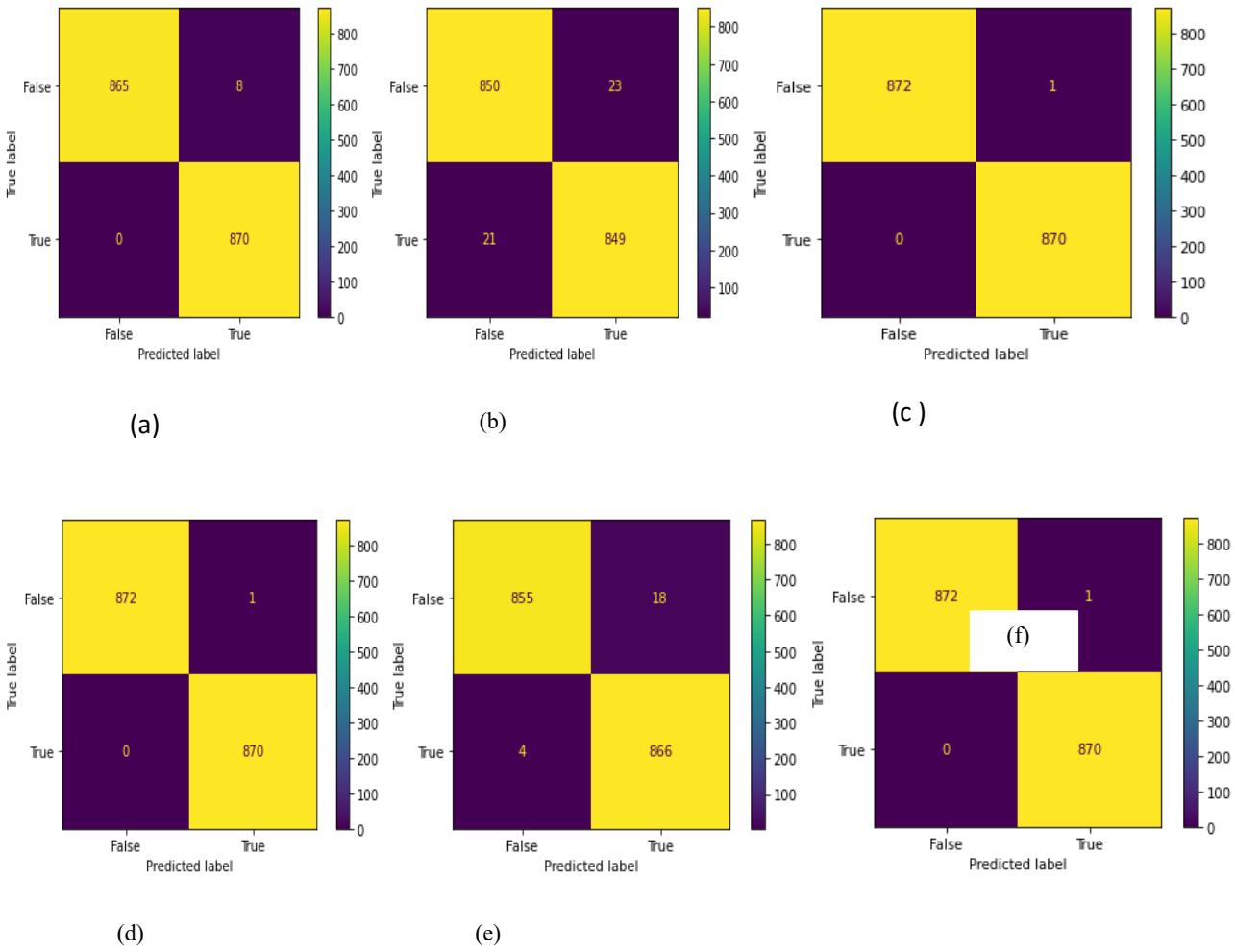


Fig.7.Confusionmatrix:(a) Logistic regression, (b) KNN classifier, (c) Decision tree classifier, (d) MLP classifier, (e) SVC classifier and (f) Random Forest classifier.

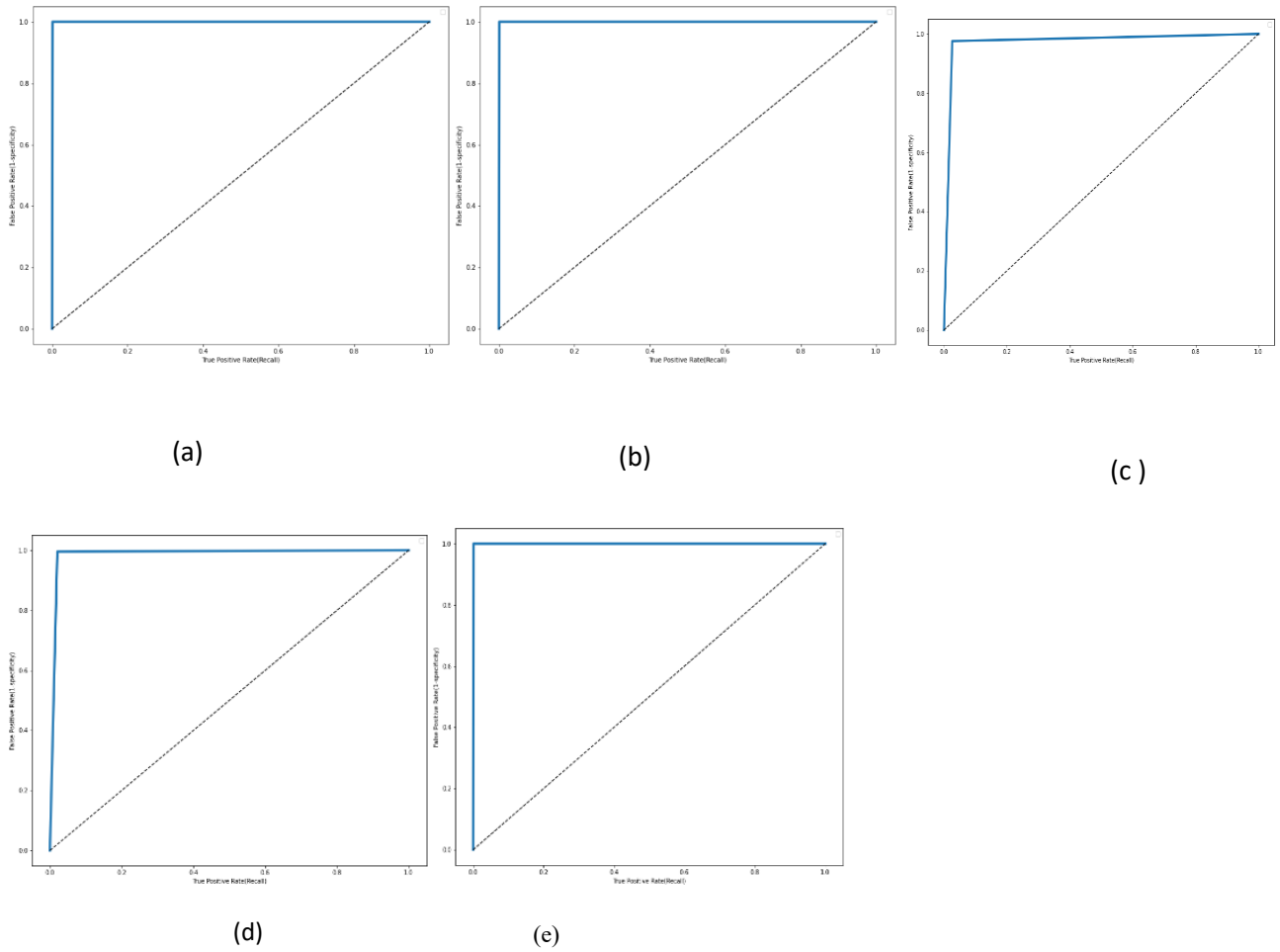


Fig.8. ROC: (a) Random Forest classifier, (b) Decision tree classifier, (c) KNN classifier, (d) ) SVC classifier, (e) MLP classifier and (f) Logistic regression.

By using the performance metrics equations (2), (3), (4),(5) and (6) on the previous confusion matrix we get this outcomes in the following table 2:

Table 2 Performance Measurement

| Algorithm | Accuracy (%) | Precision (%) | Recall (%) | Specificity (%) | F1 score (%) |
|---|---|---|---|---|---|
| **Ref[16]** | 94% | | | | |
| **Random Forest** | 99.94 | 99.88 | 1.0 | 99.88 | 99.94 |
| **Decision Tree** | 99.94 | 99.88 | 1.0 | 99.88 | 99.94 |
| **MLP** | 99.94 | 99.88 | 1.0 | 99.88 | 99.94 |
| **KNN** | 97.47 | 94.36 | 97.58 | 97.36 | 97.47 |
| **Logistic regression** | 99.54 | 99.08 | 1.0 | 99.08 | 99.54 |
| **SVC** | 98.73 | 97.96 | 99.54 | 97.93 | 98.74 |

As a result of the metrics in Table 2 above, we can observe that our prediction models' results are excellent and even exceed expectations. As observable, the Random Forest (RF), Decision Tree (DT), and Multilayer Perceptron (MLP) have recorded an excellent accuracy of 99.94%. In comparison with the accuracy of the SVC and Random Forest algorithms used in [20], we have a higher accuracy using the same dataset and the same parameters, which is a good result in the end.

## 6 Conclusion and Future work

In this work, we examine six alternative classifiers' effectiveness in predicting anemia . The experimental outcome on a test dataset indicates that Random Forest, MLP, Decision Tree, and Logistic regression perform best in terms of accuracy from KNN and SVC. Automatic prediction can lessen the amount of manual work required for diagnosis. In the future, automated technologies may be created to assist in suggesting additional diagnoses based on the predictions. These automated tools may be useful in the early identification of more serious diseases. Additionally, such a disease prediction system can be expanded to include therapy recommendations.

## 7 References


[1] Kawo, K.N., Asfaw, Z.G., Yohannes, N.: Multilevel analysis of determinants of anaemia prevalence among children aged 6–59 months in Ethiopia: classical and Bayesian approaches. Anemia 2018 (2018).

[2] Feusier, J.E.; Arunachalam, S.; Tashi, T.; Baker, M.J.; Van-Sant-Webb, C.; Ferdig, A.; Welm, B.E.; Rodriguez-Flores, J.L.; Ours, C.; Jorde, L.B.; et al. Large-scale Identification of Clonal Hematopoiesis and Mutations Recurrent in Blood Cancers. Blood Cancer Discov. 2021, 2, 226–237. [CrossRef]

[3] World Health Organization. Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity. No. WHO/NMH/NHD/MNM/11.1. World Health Organization, 2011.

[4] McLean E, Cogswell M, Egli I, Wojdyla D, De Benoist B. Worldwide prevalence of anaemia, WHO vitamin and mineral nutrition information system, 1993–2005. Public health nutrition. 2009 Apr;12(4):444-54.

[5] Prevalence of Anemia in Women of Reproductive Age, Our World in Data. Available online: https://ourworldindata.org/grapher/prevalence-of-anemia-in-women-of-reproductive-age-aged-15-29 (accessed on 28 November 2022)

[6] Khan, Jahidur Rahman, et al. "Machine learning algorithms to predict the childhood anemia in Bangladesh." Journal of Data Science 17.1 (2019): 195-218.

[7] Obaidy, Midhin AL, et al. "Prevalence and Risk Factors of Anemia among Children Aged 5 months-12 years at Al Anbar Province." Mosul Journal of Nursing 9.1 (2021): 131-137.

[8] Mattiello, Veneranda0, et al. "Diagnosis and management of iron deficiency in children with or without anemia: consensus recommendations of the SPOG Pediatric Hematology Working Group." European journal of pediatrics 179 (2020): 527-545.

[9] Meena, Kanak, et al. "Using classification techniques for statistical analysis of Anemia." Artificial Intelligence in Medicine 94 (2019): 138-152.

[10] Mukherjee, K.L., Ghosh, S., 2012. Medical laboratory Technology. Procedure Manual for Routine Diagnostic Tests. Vol I (Second edition), 263-266.

[11] Lanier, J. Brian, James J. Park, and Robert C. Callahan. "Anemia in older adults." American family physician 98.7 (2018): 437-442.

[12] Jaiswal, Manish, Anima Srivastava, and Tanveer J. Siddiqui. "Machine learning algorithms for anemia disease prediction." Recent Trends in Communication, Computing, and Electronics: Select Proceedings of IC3E 2018. Springer Singapore, 2019.

[13] Verma, Parth, and Vinay Chopra. "A Review on Machine Learning Algorithms for Anemia disease Prediction." (2022).

[14] Shilpa, S. A., Nagori, M., & Kshirsaga, V. (2011). Classification of anemia using data mining techniques. In Swarm, evolutionary, and memetic computing (pp. 113–121). Springer.

[15] El-kenawy, E.M.T. A Machine Learning Model for Hemoglobin Estimation and Anemia Classification. Int. J. Comput. Sci. Inf. Secur. 2019, 17, 100–108.

[16] Sow, B., et al.: Assessing the relative importance of social determinants of health in malaria and anemia classification based on machine learning techniques. Inform. Health Soc. Care 45(3), 229–241 (2020)

[17] Amin, N., & Habib, A. (2015). Comparison of different classification techniques using WEKA for hematological data. American Journal of Engineering Research, 4(3), 55–61.

[18] Dogan, S., Turkoglu, I.: Iron deficiency anemia detection from hematology parameters by using decision tree. International journal of Science and technology. (2008) pp: 85-92.

[19] Young, M.F.; Oaks, B.M.; Tandon, S.; Martorell, R.; Dewey, K.G.; Wendt, A. Maternal hemoglobin concentrations across pregnancy and maternal and child health: A systematic review and meta-analysis. Ann. N. Y. Acad. Sci. 2019, 1450, 47–68. [CrossRef] [PubMed]

[20] Dixit, Aditya, et al. "Prediction of Anemia Disease Using Machine Learning Algorithms." *Intelligent Computing and Networking: Proceedings of IC-ICN 2022*. Singapore: Springer Nature Singapore, 2023. 229-238.

[21] https://www.kaggle.com/datasets

[22] Kumar, Arvind, Nishant Sinha, and Arpit Bhardwaj. "A novel fitness function in genetic programming for medical data classification." *Journal of Biomedical Informatics* 112 (2020): 103623.

[23] Su, Xiaoyuan, Taghi M. Khoshgoftaar, and Russell Greiner. "Using imputation techniques to help learn accurate classifiers." *2008 20th IEEE International Conference on Tools with Artificial Intelligence*. Vol. 1. IEEE, 2008.

[24] Saidi, R., Bouaguel, W., Essoussi, N.: Hybrid feature selection method based on the genetic algorithm and pearson correlation coefficient. In: Hassanien, A.E. (ed.) Machine Learning Paradigms: Theory and Application. SCI, vol. 801, pp. 3–24. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-02357-7_1

[25] Panda, D., Dash, S.: Predictive system: Comparison of classification techniques for effective prediction of heart disease. In: Satapathy, S.C., Bhateja, V., Mohanty, J.R., Udgata, S.K. (eds.) Smart Intelligent Computing and Applications. SIST, vol. 159, pp. 203–213. Springer, Singapore (2020). https://doi.org/10.1007/978-981-13-9282-5_19

[26] . Mohan, S., Thirumalai, C., Srivastava, G.: Effective heart disease prediction using hybrid machine learning techniques. IEEE Access 7, 81542–81554 (2019)

[27] Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2(3):18–22.

[28] Díaz-Martínez, M.A., Ahumada-Cervantes, M. de los A., Melo-Morín, J.P.: Decision trees as a methodology to determine academic performance in higher education. Rev. Lasallista Investig. 18, 94–104 (2021)

[29] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning. 2002;46:389–422.

[30] Chao, C.-M., et al.: Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree. J. Med. Syst. 38(10), 1–7 (2014). https://doi.org/10.1007/s10916-014-0106-1.

[31] Sharifzadeh, F., Akbarizadeh, G., Kavian, Y.: Ship classification in SAR images using a new hybrid CNN–MLP classifier. J. Indian Soc. Remote Sens. 47(4), 551–562 (2018). https://doi.org/10.1007/s12524-018-0891-y

[32] Yeruva, Sagar, et al. "Identification of sickle cell anemia using deep neural networks." *Emerging Science Journal* 5.2 (2021): 200-21.