ANEMIA DETECTION USING MACHINE LEARNING

Chaithanya K S, Deepika M L, Harshitha M, Hemavathi S, Divyashree K, Yashodhara R

Student, Information Science and Engineering, Don Bosco Institute of Technology, Karnataka, India
Student, Information Science and Engineering, Don Bosco Institute of Technology, Karnataka, India
Student, Information Science and Engineering, Don Bosco Institute of Technology, Karnataka, India
Student, Information Science and Engineering, Don Bosco Institute of Technology, Karnataka, India
Teacher, Information Science and Engineering, Don Bosco Institute of Technology, Karnataka, India
Teacher, Information Science and Engineering, Don Bosco Institute of Technology, Karnataka, India

ABSTRACT

Computer-aided diagnosis of diseases proves to be a cost-effective solution. In addition to saving time, this approach also ensures accuracy, eliminating for additional manpower in medical decision-making processes. Various nutrition surveys indicate that nearly a quarter of the global population suffers from anemia. Therefore, there is an urgent need. to develop a proficient machine learning classifier capable of accurately detecting and classifying anemia. In this study, five ensemble learning methods - Stacking, Bagging, Voting, Adaboost, and Bayesian Boosting- are applied to four classifiers: DT, ANN, Naïve Bayes, and K-Nearest Neighbor. Theobjective is to identify which individual classifier or combination of classifiers achieves the very best accuracy in classifying blood cells for anemia detection The results demonstrate that among the ensemble methods, the stacking ensemble method attains the highest accuracy. Among the individual classifiers, the ANN performs the best while the K-Nearest Neighbor performs the worst. Interestingly, the combination of T and K-NN, when applied in the Stacking ensemble, achieves substantially better accuracy than the Artificial Neural Network alone. This highlights the reality that an ensemble of classifiers yields superior accuracy as compared to individual classifiers. Therefore, to ensure maximum accuracy in medical decision-making, anensemble of classifiers should be utilized.

Keyword: *Anemia, Hemoglobin, MCV, MCH, MCHC, Machine Learning, Logistic regression, K Nearest Classifier, Random Forest, Decision Tree, Artificial neural network.*

1. INTRODUCTION

Anemia occurs when the quantity of hemoglobin in someone's blood drops under normal levels. A decrease in hemoglobin is proportional to a decrease in RBC count. RBCs contain hemoglobin which aids in the transportation of oxygen from the lungs to the various elements of the body[1][5]. Insufficient supply of oxygen to many tissues and organs can have a negative impact on the body. Individuals affected by anemia may experience fatigue, weakness, and a lack of strength[3]. Anemia also can affect folks who do not have proper and regular diet rich in iron and vitamins. Additionally, people tormented by continual diseases such as kidney disease, diabetes, cancer, a family history of inherited anemia, chronic infections such as tuberculosis or HIV, and those who've had sizable blood loss from injury or surgery are probably to be vulnerable. Anemia may be labeled into various ranges consisting of mild, moderate, or severe, depending on how much the RBC count and/or hemoglobin levels are decreased[7][8]. The main causes of anemia include a decreased production of RBCs by the bone marrow due to deficiencies in nutrition such as iron and vitamin B, bone marrow failure, or diseases such as infection in the bone marrow, lymphoma, solid tumor, etc. a few researchers have used rare bureaucracy and calculating depictions of some pink blood cells to identify[1][4].Anemia caused by a lack of iron the use of three different special classifiers. However, they failed to anticipate the cause for such issues when trying to categories two kinds of anemia[6][5].The Hemoglobin checks the quantity of hemoglobin within the blood. If it's far observed that the hemoglobin level is lower than normal, it means the person has a low crimson blood cell relay (anemia). The normal range for hemoglobin is: [Specify the normal range for hemoglobin[3][6].

This paper explores the application of five popular ensemble methods (Bagging, Adaboost, Stacking, Bayesian Boosting, and Voting) on four base learners (Decision Tree, ANN, Naïve Bayes, and K-Nearest Neighbor) to detect anemia[1].What units our research aside is that we've advanced our own software program especially for anemia detection, and the dataset used in this examine changed into prepared the usage of our software program. the principal goal of this research is to become aware of the first-rate aggregate of classifiers for Voting and Stacking techniques in accurately classifying red blood cells (RBC)[4][7]. In addition, we intention to determine the best classifier among DT, ANN, Naïve Bayes, and K-NN for Adaboost, Bayesian Boosting, and Bagging techniques. Finally, a comprehensive contrasting the five ensemble learning techniques (Bagging, Adaboost, Stacking, Bayesian Boosting, and Voting) primarily according to the information provided four classifiers (DT, ANN, Naïve Bayes, and K-NN) for RBC classification will be conducted[5][3].

2. LITERATURE SURVEY

1. Simulation Model for Anemia Detection using RBC counting algorithm digital photograph processing is in recent times broadly utilized in the field of biomedicine programs. Counting of the pink blood cells pictures is useful for detecting anemia because the guide place identity and counting of pink blood cells is a tedious, mistakes prone and time ingesting there's a growing want for automating the complete process. A simulation version to detect anemia, RBC counting is utilized set of rules is supplied in this paper, both circular Hough rework and connected issue Labelling are applied for counting the quantity of RBCs and the outcomes are as compared. Watershed rework Separating overlapping blood cells is done likewise parameters such as segmentation accuracy, sensitivity, and specificity also are covered on this paper.

2. Machine Learning Algorithms for Anemia disease Prediction the remarkable advances in health industry have led to a significant production of data in everyday life. This data requires processing to extract use- full information, which can be beneficial for analysis, prediction, recommendations, and decision-making records mining and system studying strategies are used to transform the available facts into treasured information. In medical science, disease prediction at the right time is the central problem for professionals for prevention and effective treatment plan. Sometimes, in absence of accuracy this may lead to death. This study investigates algorithms for supervised machine learning - NB, RF, and DT algorithm in order to predict anemia using CBC (complete blood count) data collected from pathology centers. Results indicate that Naive-Bayes technique out plays in phrases of accuracy in comparison to C4.5 and random forest.

3. Prediction of Anaemia among children using Machine Learning anemia is a first-rate public fitness problem in particular common place amongst preschool-elderly youngsters and ladies in maximum in the process of developing international locations. It is extra exciting to understand the possibilities of anemia given the related elements, instead of know-how the factors on my own. So, this look at geared toward constructing a few predictive fashions utilizing usage of the diagnosed risk factors thru device mastering technique. A health center primarily based pass-sectional examine became performed. The members in this observe included youngsters of age organization 6-36months. We advanced a few ML algorithms which includes linear Anemia Detection using device mastering ,CART, KNN, random woodland and logistic regression (LR) if you want to are expecting the anemia popularity of youngsters beneath 36 months antique in Jammu. We in comparison the anticipations acquired with the consequences of LR that's the maximum broadly classifier in prediction disorder reputation. It turned into observed that the LR classifier confirmed an the accuracy of predicting the anemia is 61.67%. The opposite predictive fashions showed up by accident nearly comparable detection accuracy because the logistic regression. additionally, we've determined that k-NN received the less accuracy in appearing predictions however random woodland confirmed the quality all accuracy the predictions is 67.18%. fashions constructed using system studying strategies. Our look at also states that ML-A are capable of predicting anemia based totally on the commonplace threat factors which in the long run can be beneficial to save you and manipulate early life anemia.

4. Comparative Study Between Decision tree, SVM and KNN to Predict Anaemic Condition Anemia, a disorder that is because of an inadequacy of hemoglobin or crimson blood cells in the blood. It is very risky on the time of

pregnancy, menstruation and in ICU sometimes causing death. So, it's a need hemoglobin and detects anemia quickly. Usually, doctors examine the eye conjunctiva color and confirmed by a blood test that hurts, time-consuming and costly. In this observe, total 104 people (54 males and 50 females) is collected with their clinical blood hemoglobin level, anemic condition and taken palpebral conjunctiva image. The images are captured with a cell phone camera of good resolution. By using the images, the percentage of the red, green and blue pixels are extracted in MATLAB, image processing method. Taking those features, the Hemoglobin level is plotted. In total 81 data is taken for training purposes and 23 data for testing. For Anemia detection, the 81 data are trained with a used different classifier like as Linear SVM, Coarse Tree, and Cosine KNN and have been got maximuaccuracy of 82.61% in Decision Tree (Coarse) by testing 23 data.

3. METHODOLOGY

On this paper we've got used 4 system getting to know algorithms to expect anemia in patients according to six attributes for 2 special instructions over records set containing 1422 data obtained from laboratories. Here a short overview of the methodology, the principles behind it, and characteristics are mentioned.

Data collection:

We have considered online data sets which have a record of 1421 patient data having 6 attributes namely gender, hemoglobin, MCH, MCHC, MCV and result.

Data Pre- Processing:

In this step we clean the data by checking for null and duplicated values as these may tend to over-fit the model. We also implement transformation technique to make the facts within the more comprehensible form.

Exploratory Data Analysis:

This records evaluation method will usefully resource in visually determining the connection between the attributes through plotting them the usage of various plotting techniques including container and whisker's plot, scatter plot and many others.

Modelling

The statistics is split into classes namely train data and test facts that have a ratio of 7:3 respectively. We teach the version using four type algorithms namely Logistic Regression, choice tree, Random woodland, k Nearest Classifier.

3.1 Types of ML Classifiers

1) Decision Trees: Decision Tree classifiers utilize the tree structure to represent the dataset. This method is specifically used for determining discrete valued target functions. Instances are categorized by traversing down the tree from the root to a leaf node. Every node in the tree indicates an attribute, while every department represents a value associated with that attribute.

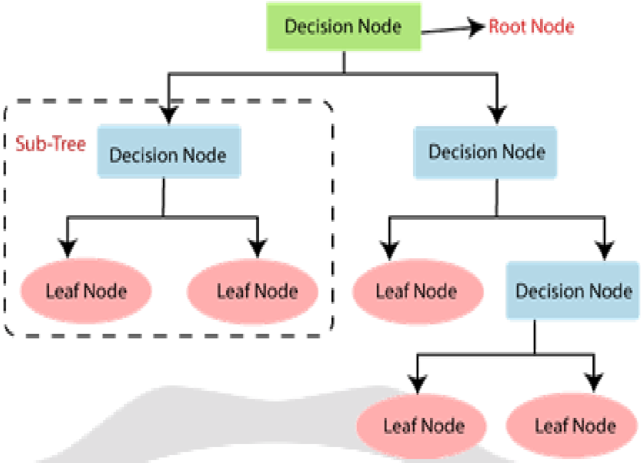


Fig-1: Decision Tree

2) Random forest: Random Forest is a supervised learning algorithm. The "forest" it builds, is a group of decision trees which is generally trained with the "begging" methodology. The idea of the bagging method is that a combinationof learning models increases the increases result.

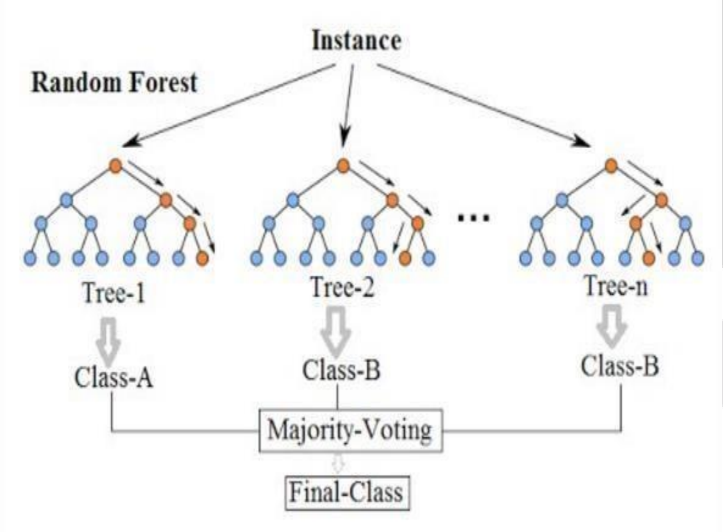


Fig-2: Random Forest

3) Naïve Bayes Classifier: It classifier makes use of independent assumptions and Bayes Theorem. It operates under the hypothesis that the presence or the absence of selected attribute is not correlated with the presence or absence of other attributes. This classifier requires minimal training data to calculate the suggest and variability of the associated variables.

4) K-Nearest Neighbor Classifier: The K-NN algorithm classifies data by comparing testing examples with similar training examples. In K-NN, k represents a small positive integer. To classify an unseen example, the majority vote of the neighbors is considered. If k equals 1, the example is assigned to the class Classifier

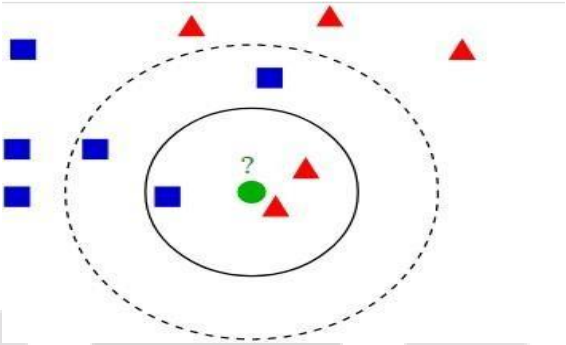


Fig-3 : K-Nearest Neighbor Classifier

5) Logistic Regression: Logistic Regression is also be called Linear Regression model, but the Logistic Regression uses a more complex function, this function can be defined as the 'Sigmoid function' or also called as the 'logistic function' instead of a liner function. The theory of logistic regression limits the function between the range 0 and 1. Hence liner functions fail to be represented as it can have a value is more than 1 or less than 0which is which as per the theory of logistic regression. LR is dented by the following expression:0≤ f(x) ≤ 1.

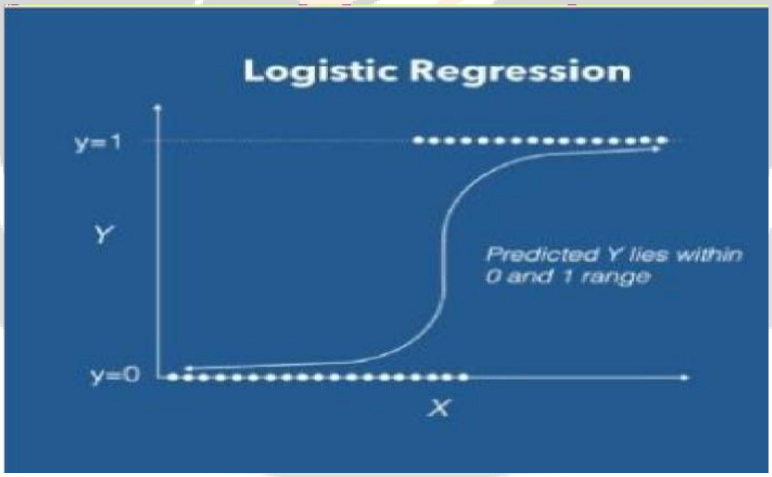
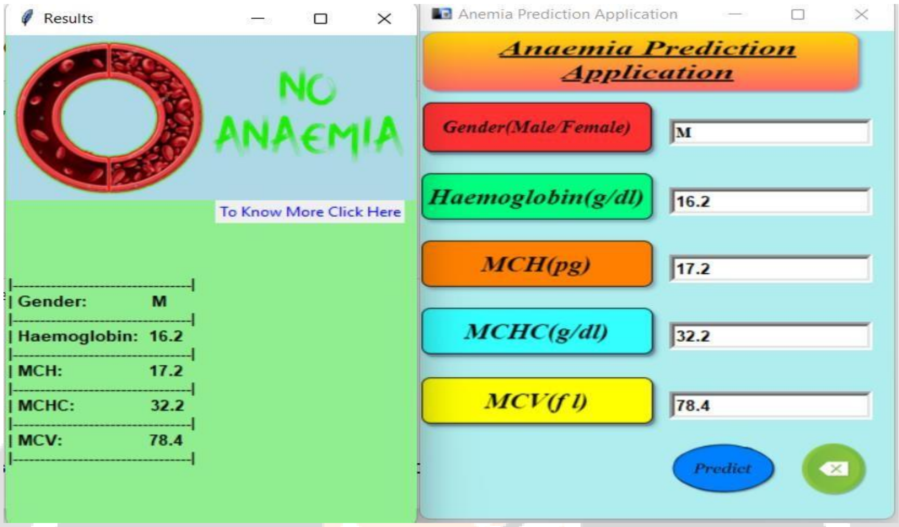


Fig-4: Logistic Regression

4. CONCLUSION

Reduction of RBC's causes deficiency of oxygen, which drives the values of hemoglobin, MCH, MCHC, MCV to be out of their normal stages which might also lead to severe ailments if no longer identified at an early degree. In this project, 4 different algorithms (LR, KNN, DT, RF) have been implemented to discover anemia below the attention of six attributes using 1421 samples. It is seen that LR showed better overall performance compared to other algorithms with accuracy of 95.32%. The above model is refined using various model evaluation strategies to

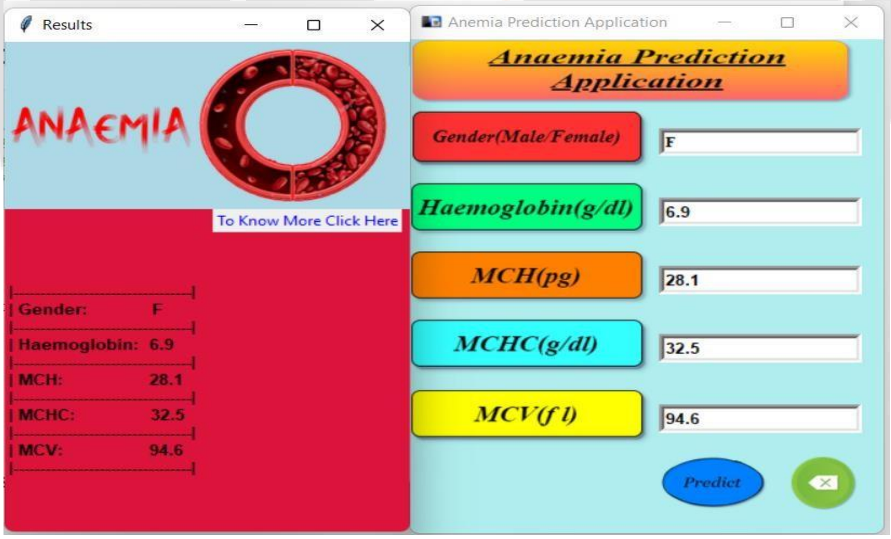
determine the risk factors that influence the prediction system. Hence, we conclude that, LR have low well-known deviation which means that maximum of the numbers is close to the imply value. The model is then deployed into a standalone application which it there are various applications for this in as a reference in diagnostic centers for anaemic prediction.
The below image shows the graphical user interface where the values are entered and the result is displayed as shown.



```
|------------------------|
| Gender:          M
|------------------------|
| Haemoglobin:  16.2
|------------------------|
| MCH:             17.2
|------------------------|
| MCHC:            32.2
|------------------------|
| MCV:             78.4
|------------------------|
```

Gender(Male/Female): M
Haemoglobin(g/dl): 16.2
MCH(pg): 17.2
MCHC(g/dl): 32.2
MCV(f l): 78.4

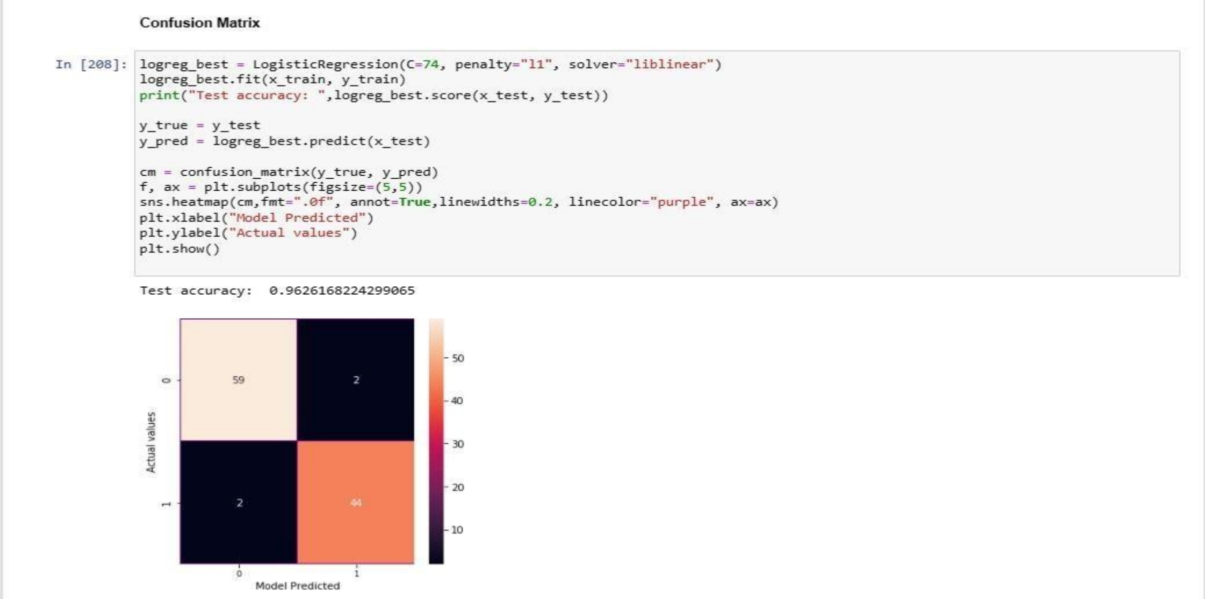
Fig-5: Application Predicted for Non-Anemic patient



```
|------------------------|
| Gender:          F
|------------------------|
| Haemoglobin:  6.9
|------------------------|
| MCH:             28.1
|------------------------|
| MCHC:            32.5
|------------------------|
| MCV:             94.6
|------------------------|
```

Gender(Male/Female): F
Haemoglobin(g/dl): 6.9
MCH(pg): 28.1
MCHC(g/dl): 32.5
MCV(f l): 94.6

Fig-6: Application Predicted for Anemic patient



Confusion Matrix

```
In [208]: logreg_best = LogisticRegression(C=74, penalty="l1", solver="liblinear")
          logreg_best.fit(x_train, y_train)
          print("Test accuracy: ",logreg_best.score(x_test, y_test))

          y_true = y_test
          y_pred = logreg_best.predict(x_test)

          cm = confusion_matrix(y_true, y_pred)
          f, ax = plt.subplots(figsize=(5,5))
          sns.heatmap(cm,fmt=".0f", annot=True,linewidths=0.2, linecolor="purple", ax=ax)
          plt.xlabel("Model Predicted")
          plt.ylabel("Actual values")
          plt.show()

          Test accuracy:  0.9626168224299065
```

Fig-7: Confusion Matric

5. REFERENCES

[1] P. Rakshit, “Detection of Abnormal Findings in Human RBC in Diagnosing G -6-P-D Deficiency Hemolytic Anaemia Using Image Processing,” 2013 IEEE 1st International Conference on Condition Assessment Techniquesin Electrical Systems (CATCON), pp. 297– 302, 2013.

[2] Khan, R. K. Mondol, M. A. Zamee, and T. Tarique, “Regression Model Based On FPGA,” 2014 International Conference on Informatics, lectronics & Vision (ICIEV), pp. 1-5, 2014.

[3] Rahul Joshi and Minyechil Alehegn, “Analysis and Prediction of diabetes diseases using machine learning algorithm”: Ensemble approach, International Research Journal of Engineering and Technology Volume: 04 Issue:10 | Oct -2017

[4] M. Tyagi, L. M. Saini, and N. Dahyia, “Detection of Poikilocyte Cells in Iron Deficiency Anaemia Using Artificial Neural Network,” 2016 International Conference on Computation of Power, Energy Information and Commuincation (ICCPEIC), pp. 108–12, 2016.

[5] Dr. M. Renuka Devi and J. Maria Shyla, “Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus”, International Journal of Applied Engineering Research ISSN 0973 - 4562 Volume 11, Number 1 (2016)pp 727-730 © Research India.

[6] S. Belginova and I. Uvaliyeva, “Decision Support System for Diagnosing Anemia,” 2018 4th International Conference on Computer and Technology Applications (ICCTA), no. Mcv, pp. 211– 215, 2018

[7] Zhilbert Tafa and Nerxhivan Pervetica, “An Intelligent System for Diabetes Prediction”, 4th Mediterranean Conference on Embedded Computing MECO – 2015 Budva, Montene

[8] Deeraj Shetty, Kishor Rit, Sohail Shaikh and Nikita Patil “Diabetes Disease Prediction Using Data Mining”.International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS) 2016.