e-ISSN: 2582-5208

International Research Journal of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

MACHINE LEARNING ALGORITHMS FOR ANEMIA DISEASE PREDICTION - A REVIEW

Parth Verma[*1], **Dr. Vinay Chopra**[*2]

[*1]Student, Department Of Computer Science And Engineering, DAV Institute Of Engineering And Technology, Jalandhar, Punjab, India.

[*2]Professor, Department Of Computer Science And Engineering, DAV Institute Of Engineering And Technology, Jalandhar, Punjab, India.

ABSTRACT

Remarkable advances in the healthcare industry are generating important data in our daily lives. This data needs to be processed to extract useful information that may be useful for analysis, prediction, recommendations, and decision making. Transform available data into valuable information using data mining and machine learning techniques. In medicine, timely disease prediction is a central issue for professionals for prevention and effective treatment planning. From time to time, a lack of accuracy can be fatal. This study examines monitored simple Bayes, random forest, and decision tree machine learning algorithms for predicting anemia using CBC (Complete Blood Count) data collected from the Pathology Center. The results show that the Naive Bayes method is superior in accuracy compared to C4.5 and Random Forest.

Keywords: Anemia, Classification Algorithms, Complete Blood Count (CBC), Decision Making.

I. INTRODUCTION

Modern medical systems generate vast amounts of data every day. This data needs to be mined and analyzed to extract useful information and reveal hidden patterns. Data mining is the process of discovering new patterns in data collected from various sources. Many machine learning algorithms are used for forecasting in various areas such as healthcare, weather forecasting, stock price forecasting, and product recommendations. An important aspect of medical research is the prediction of various illnesses and the factors that cause them. In the medical field, health data is used to predict epidemics, detect illnesses, improve quality of life, and prevent premature death [1]. In this task, we will look at three different classification algorithms for prediction.

Anemia is defined as a decrease in the number of red blood cells (RBCs) or hemoglobin in the blood [2], which not only adversely affects economic and social development, but also causes serious adverse health effects. The most reliable indicator of anemia is hemoglobin concentration in the blood, but there are several factors that can cause anemia. B. Iron deficiency, HIV, malaria, chronic infections such as tuberculosis, vitamin deficiency, z. Vitamin B12 and A, cancers, and acquired diseases that affect red blood cell production and hemoglobin synthesis.

Anemia causes fatigue and decreased productivity [3, 4, 5], and when it occurs during pregnancy, it can be associated with an increased risk of maternal and perinatal mortality [6, 7]. According to the World Health Organization (WHO), the mortality rate of mothers and newborns in developing countries in 2013 was 3 million. Prediction of anemia disease plays a very important role in detecting other related diseases. Anemia disorders are categorized based on morphology or root cause (Figure 1). Anemia is divided into three types, normocytic, microcytic, and macrocytic, based on morphology. Anemia is classified into three types depending on the cause: blood loss, insufficient production of normal blood, and excessive destruction of blood cells.

In this article, we will use a dataset collected from a local pathology center to examine the performance of Naive Bayes, Random Forest, and decision tree algorithms for predicting anemic disease. The need for this investigation stems from the fact that the root cause of the disease varies from region to region. Random forest classifiers have been previously studied to predict heart disease and chronic kidney disease, but as far as we know, they have not been studied to predict anemia. This adds novelty to the work.

The rest of the paper is structured as follows: Section 2 outlines existing related work. Section 3 describes different types of anemia diagnostic tests. Section 4 shows the proposed methodology. Section 5 provides details and a discussion of the experiment. Eventually, we end up in Section 6.

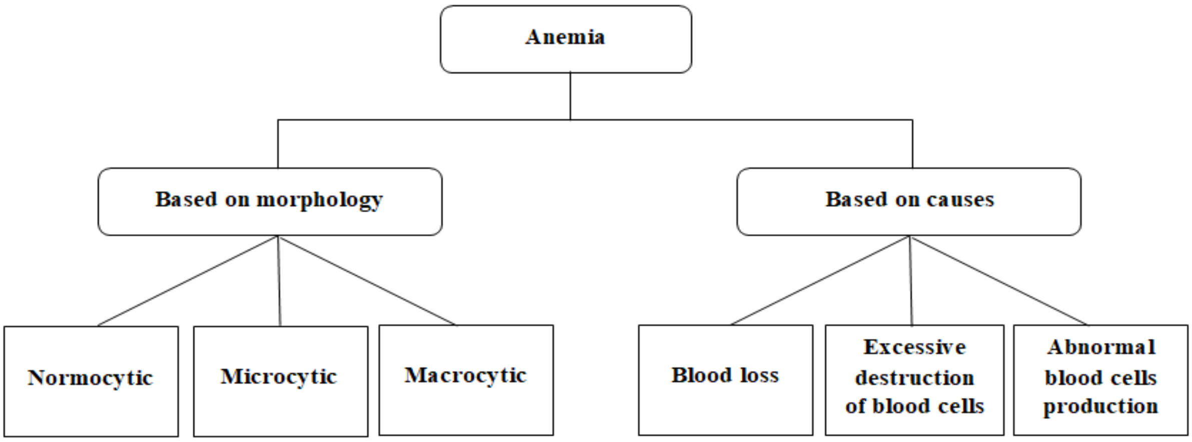

Figure 1: Classification of Anemia.

II. RELATED WORKS

Numerous data mining and machine learning techniques for anemic diseases have been used over the last decade. The most frequently mentioned are:

In [8], to predict anemia, the author has used the SMO support vector machine and the C4.5 decision tree algorithm to compare the performance of the two algorithms.

In [11], WEKA was used to obtain classifiers suitable for developing mobile apps that can predict and diagnose comments on blood data. The authors compared neural network classification algorithms using the J48 and a naive Bayes classifier. The results show that the J48 classifier has the highest accuracy.

Dogan & Turkoglu [12] has developed a decision support system to detect iron deficiency anemia using a decision tree algorithm. This algorithm uses three hematological parameters: serum iron, serum iron-binding capacity, and ferritin. The assessment was based on data from 96 patients and the results were well compared to the physician's decision.

Abdullah and Alasmari [13] experimented with the WEKA algorithm (Naive Bayes, Multilayer Perception, J48, and SMO) to predict the type of anemia from the CBC report. The analysis was based on actual data generated from CBC reports from 41 anemia patients. Similar to [11], the J48 decision tree algorithm and SMO performed best with 93.75% accuracy.

In contrast to the work in [11] and [13], we have chosen a different set of classifiers and local data.

Diagnostic tests Classification

There are four major tests ordered to diagnose anemia disorders: complete blood count (CBC), ferritin, PCR (polymerase chain reaction), and hemoglobin electrophoresis.

- The CBC test is the most common blood test used to measure general health and diagnose various illnesses [8] such as anemia, infections, and leukemia. The complete blood cell count test measures about 15 tests, including hemoglobin (Hb), red blood cells (RBC), hematocrit (HCT), mean corpuscular hemoglobin (MCH) and mean corpuscular volume (MCV) [8].
- The ferritin test measures the amount of iron stored in the body. High ferritin levels indicate impaired iron storage such as hemochromatosis. Low ferritin levels indicate iron deficiency that causes anemia.
- The PCR test is a molecular test used to diagnose hereditary diseases.
- A hemoglobin electrophoresis test is a blood test used to measure and identify different types of hemoglobin in the bloodstream.

III. METHODOLOGY

We used three classifiers: Random Forest, Naive-Bayes, and Decision Tree C4.5 algorithm. Figure 2 shows a flow chart of the proposed procedure.

Random Forest Algorithm

The Random Forest (RF) algorithm is derived from the decision tree classifier. This is a combination of tree predictors that aggregates the results of all the trees in the collection and uses a majority vote in the prediction.

Decision Tree Algorithm

A decision tree is a tree in which each branch node represents a choice from multiple choices and each leaf node represents a decision. Widely used in various fields [9] [10]. C4.5 (WEKA J48) is a decision tree developed by Ross Quinlan.

Naive – Bayes Algorithm

Naive-Bayes algorithm based on Bayes rules for conditional probabilities. Get all the attributes contained in the data and analyze them individually as if they were equally important and independent of each other. In this, a very small amount of training data is required.

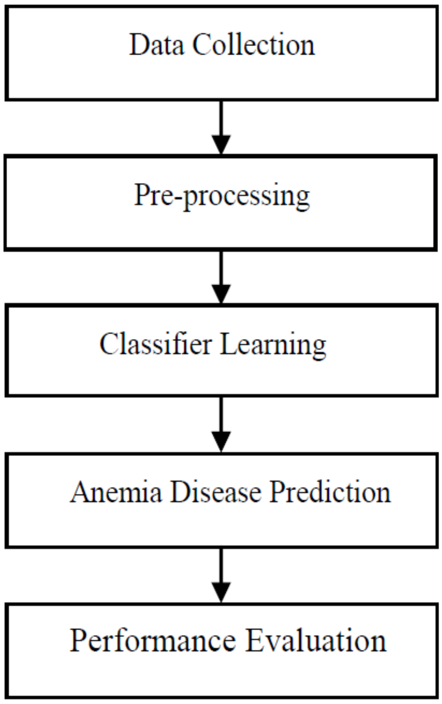

Figure 2: Flowchart of proposed Model.

IV. RESULTS AND DISCUSSION

Dataset

We collect data from various nearby pathology and laboratory centers. The data set collected consists of 200 test patterns. This is CBC test data. The dataset contained 18 attributes, and we selected only the attributes needed to detect anemic disease. These are age, gender, MCV, HCT, HGB, MCHC, and RDW.

Experimental setup

The proposed method uses CBC test values. First, as described in 5.1, we preprocess the data and extract seven attributes. Then apply a random forest, decision tree, and NB classifiers. Performance is evaluated in terms of accuracy and mean absolute error (MAE). Mean absolute error (MAE) measures how close a prediction is to the final result. Table 1 shows the results of the three classifiers and 10-fold cross-validation was used for accuracy.

Table 1. Comparison of Algorithms.

SN.	Classifier	Mean-Absolute Error	Accuracy
1	Random Forest	0.0332	95.3241
2	Naïve-Bayes	0.0333	96.0909
3	C4.5	0.0347	95.4502

Figure 3 and Figure 4 show a comparison of the accuracy and performance of each classification algorithm based on MAE. The Naive Bayes classifier shows the best performance in the dataset, as opposed to [11] and [13]. This is not surprising, as the datasets used in these studies are diverse and the etiology of the disease can vary from country to country. We achieve maximum accuracy of 96.09% with the NB classifier. This is superior to the best performing classifiers SMO and J48 with 93.75% accuracy reported in [13].

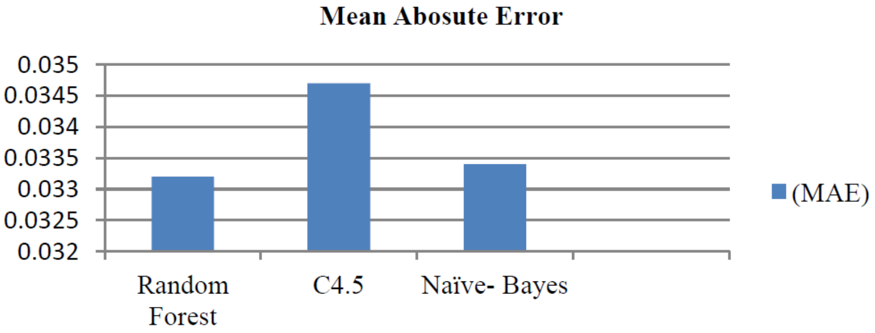

Figure 3: MAE using each Algorithm.

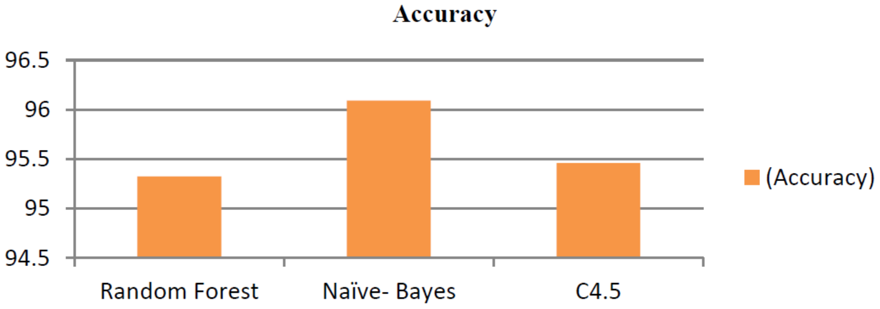

Figure 4: Comparison of accuracy using each Algorithm.

V. CONCLUSION

In this article, we compared the performance of three different classifiers in predicting anemia. Experimental results from the sample dataset suggest that the Naive Bayes classification algorithm has the best performance in terms of accuracy compared to C4.5 and Random Forest. Automatic prediction can reduce the manual labor involved in diagnosis. In the future, we can develop automated tools that can support forecast results and suggest further diagnostics. Such automated tools help detect more serious illnesses in a timely manner. In addition, such disease prediction systems can be extended to recommend treatment plans.

VI. REFERENCES

[1] Arun, V, et al.: Privacy of Health Information in Telemedicine on Private Cloud, International Journal of Family Medicine & Medical Science Research. (2015)

[2] Provenzano, R., Lerma, E.V., & Szczech, L.: Management of Anemia. Springer.(2018)

[3] Ezzati, M., Lopez, Ad., Rodgers, A., Murray, C.J.L.: Comparative quantification of health risks: global and regional burden of disease attributable to selected major risk factors. Geneva: World Health Organization. (2004)

[4] Balarajan, Y., et al.: Anaemia in low-income and middle-income countries. (2011)

[5] Haas, J.D., Brownlie, T.: Iron deficiency and reduced work capacity: A critical review of the research to determine a causal relationship. J Nutr. (2001)

[6] Kozuki, N., Lee, A.C., Katz, J.: Child Health Epidemiology Reference Group. Moderate to severe, but not mild, maternal anemia is associated with increased risk of small-for-gestational-age outcomes. J Nutr. (2012)

[7] Steer, P.J.: Maternal hemoglobin concentration and birth weight. Clin Nutr. (2000)

[8] Shilpa A. Sanap, Meghana Nagori, Vivek Kshirsaga.: Classification of Anemia Using Data Mining Techniques.: Swarm, Evolutionary, and Memetic Computing pp 113-121. Springer (2011).

[9] Jerez-Aragonés J.M. et al.: A combined neural network and decision trees model for prognosis of breast cancer relapse. Artif Intell Med. (2003) pp 45–63.

[10] Podgorelec, V. et al.: Decision trees: an overview and their use in medicine. J Med Syst. (2002) pp: 445–463

[11] N. Amin and A. Habib Comparison of different classification techniques using WEKA for hematological data, American Journal of Engineering Research, Volume-4, Issue-3, pp-55-61 (2015)

[12] 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.

[13] Manal Abdullah and Salma Al-Asmari, Anemia types prediction based on data mining classification algorithms, Communication, Management and Information Technology, (2016) Taylor & Francis Group, London,