International Research Journal of Education and Technology
ISSN 2581-7795

A Review on Machine Learning Algorithms for Anemia disease Prediction

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

[1]*Student, Dept. of Computer Science and Engineering, DAV Institute of Engineering and Technology, India*
[2] *Professor, Dept. of Computer Science and Engineering, DAV Institute of Engineering and Technology, India*

Abstract - *In our daily lives, remarkable advancements in the healthcare industry are creating essential data. This data must be processed in order to extract valuable information for analysis, forecasting, providing suggestions, and making decisions. Using data mining and machine learning approaches turn existing data into useful knowledge. In medicine, accurate illness prediction is critical for both preventive and effective treatment planning. A lack of accuracy might be fatal on occasion. Using CBC (Complete Blood Count) data from the Pathology Center, this study investigates monitored basic Bayes, random forest, and decision tree machine learning algorithms for predicting anemia. The results reveal that the Naive Bayes technique outperforms C4.5 and Random Forest in terms of accuracy.*

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

1. INTRODUCTION

Every day, modern medical systems create massive amounts of data. To extract relevant information and expose hidden patterns, this data must be mined and evaluated. The practice of uncovering new patterns in data obtained from diverse sources is known as data mining. In fields as diverse as healthcare, weather forecasting, stock price forecasting, and product recommendations, a variety of machine learning techniques are utilized for forecasting. Prediction of various illnesses and the elements that cause them is an essential aspect of medical research. Health data is used to anticipate epidemics, identify illnesses, improve quality of life, and avoid early mortality in the medical industry [1]. We'll look at three alternative categorization methods for prediction in this challenge.

Anemia is defined as a decrease in the amount of red blood cells (RBCs) or hemoglobin in the blood [2], which has major health consequences as well as a negative impact on economic and social growth. The concentration of hemoglobin in the blood is the most reliable sign of anemia; however anemia can be caused by a variety of reasons. B. Iron deficiency, HIV, malaria, tuberculosis, vitamin deficit, z. Vitamin B12 and A deficiency, malignancies, and acquired disorders that impact red blood cell development and hemoglobin synthesis.

Anemia causes fatigue and decreased productivity [3, 4, 5], and it has been linked to an increased risk of maternal and perinatal death [6, 7] when it occurs during pregnancy. In 2013, the World Health Organization (WHO) estimated that 3 million mothers and babies died in poor nations. The ability to predict anemia sickness is crucial in diagnosing other diseases that are linked to it. Anemia disorders are classified according to their morphology or root cause (Figure 1). Based on appearance, anemia is classified as normocytic, microcytic, or macrocytic. Anemia is divided into three forms based on the cause: blood loss, insufficient normal blood production, and excessive blood cell death.

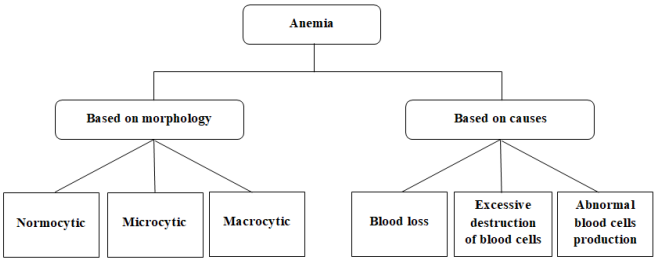

Fig -1: Classification of Anemia

In this article, we'll look at the performance of Naive Bayes, Random Forest, and decision tree algorithms for predicting anemia using data from a local pathology clinic. The need for this inquiry derives from the fact that the disease's core cause differs by place. Random forest classifiers have previously been used to predict heart disease and chronic renal disease, but they have not been used to predict anemia as far as we know. This adds to the work's uniqueness.

The following is how the rest of the paper is organized: The section 2 summarizes similar work that has already been done. The section 3 discusses the various forms of anemia diagnostic testing. The proposed methodology is shown in section 4. The experiment is described in full and discussed in section 5. We eventually get to section 6.

2. RELATED WORK

Over the recent decade, a variety of data mining and machine learning strategies for anemic disorders have been used. The following are the most frequently mentioned:

The author of [8] employed the SMO support vector machine and the C4.5 decision tree method to examine the performance of the two systems in predicting anemia.

WEKA was utilized in [11] to generate classifiers suited for constructing mobile apps that can predict and diagnose blood data comments. The authors used the J48 and a naïve Bayes classifier to compare neural network classification techniques. The J48 classifier has the highest accuracy, according to the data.

Using a decision tree algorithm, Dogan & Turkoglu [12] created a decision support system to identify iron deficiency anemia. Three hematological parameters are used in this algorithm: serum iron, serum iron-binding capacity, and ferritin. The results were well compared to the physician's choice, and the assessment was based on data from 96 patients.

Abdullah and Alasmari [13] used the WEKA algorithm to predict the type of anemia from the CBC report (Naive Bayes, Multilayer Perception, J48, and SMO). The study was based on real-world data from 41 anemia patients' CBC findings. The J48 decision tree algorithm with SMO performed best with 93.75 percent accuracy, similar to [11].

We used a different collection of classifiers and local data than the authors of [11] and [13].

2.1 Diagnostic tests Classification:

Complete blood count (CBC), ferritin, PCR (polymerase chain reaction), and hemoglobin electrophoresis are the four major tests used to diagnose anemia problems.

- The CBC test is the most frequent blood test used to assess overall health and diagnosis conditions such as anemia, infections, and leukemia [8]. Hemoglobin (Hb), red blood cells (RBC), hematocrit (HCT), mean corpuscular hemoglobin (MCH), and mean corpuscular volume (MCV) are among the 15 assays measured by the complete blood cell count [8].
- The ferritin test determines how much iron is stored in the body. Iron storage problems, such as hemochromatosis, are indicated by high ferritin levels. Low ferritin levels indicate anemia due to iron deficiency.
- A molecular test used to determine genetic illnesses is the PCR test.
- A hemoglobin electrophoresis test is a blood test that measures and distinguishes between different forms of hemoglobin in the blood.

3. METHODOLOGY

Random Forest, Naive-Bayes, and the Decision Tree C4.5 algorithm were employed as classifiers. The proposed approach is depicted in Figure 2 as a flow chart.

3.1 Random Forests:

The decision tree classifier gives rise to the Random Forest (RF) method. This is a set of tree predictors that combines the findings of all the trees in the collection and makes a prediction based on a majority vote.

3.2 Decision Tree:

A decision tree is a tree in which each leaf node represents a decision and each branch node represents a choice among numerous options [9] [10]. It's widely used in a variety of disciplines. Ross Quinlan created the C4.5 (WEKA J48) decision tree.

3.3 Naïve-Bayes Algorithm:

It is a simple algorithm that can be used to solve a variety of problems. For conditional probabilities, the Naive-Bayes method uses Bayes principles. Gather all of the data's properties and examine each one separately as though they were equally relevant and unrelated. Only a modest amount of training data is needed in this case.

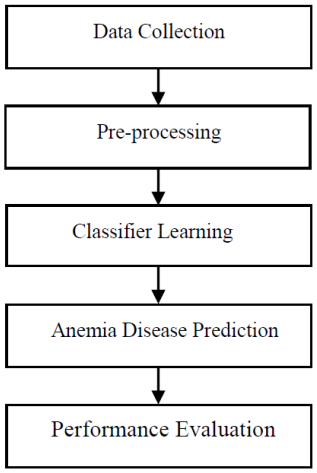

Fig -2: Flowchart of proposed Model.

4. RESULT AND DISCUSSION

4.1 Dataset:

We gather information from a number of pathology and laboratory centers in the area. A total of 200 test patterns were recorded in the data set. This is the result of a CBC test. We chose only the attributes needed to detect anemic illness from the dataset's 18 attributes. Age, gender, MCV, HCT, HGB, MCHC, and RDW are the variables.

4.2 Experimental Setup:

CBC test values are used in the suggested technique. We begin by preprocessing the data and extracting seven attributes, as stated in Section 5.1. Then use a random forest, decision tree, and NB classifiers to refine your results. Accuracy and mean absolute error are used to assess performance (MAE). The mean absolute error (MAE) is a metric that quantifies how near a prediction is to the actual outcome. The results of the three classifiers are shown in Table 1, and accuracy was determined using 10-fold cross-validation.

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

Figures 3 and 4 depict a comparison of the accuracy and performance of each MAE-based classification algorithm. In comparison to [11] and [13], the Naive Bayes classifier performs the best in the dataset. This is unsurprising given the variety of datasets utilized in these studies and the fact that the etiology of the disease varies per nation. The NB classifier achieves a maximum accuracy of 96.09 percent. This outperforms the best-performing classifiers, SMO and J48, which have a claimed accuracy of 93.75 percent in [13].

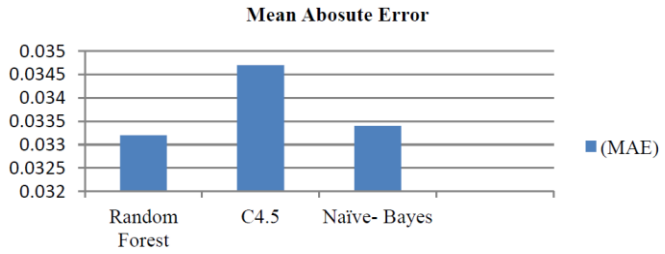

Fig -3: MAE using each Algorithm.

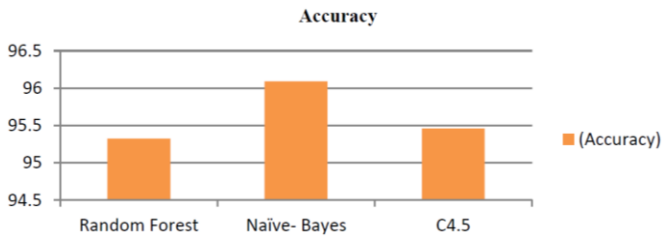

Fig -4: Comparison of accuracy using each Algorithm.

5. CONCLUSION

We compared the performance of three different classifiers in predicting anemia in this article. In comparison to C4.5 and Random Forest, experimental results from the sample dataset reveal that the Naive Bayes classification method has the best accuracy. The manual work involved in diagnosis can be reduced by using automatic prediction. We can create automated tools in the future to help forecast findings and propose more diagnostics. Such automated methods aid in the early detection of more serious ailments. Furthermore, such disease prediction algorithms can be used to make therapy recommendations.

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.: Anemia 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,