Machine Learning Algorithms for Anemia Disease Prediction

Manish Jaiswal, Anima Srivastava and Tanveer J. Siddiqui

Abstract The remarkable advances in health industry have led to a significant production of data in everyday life. These data require processing to extract useful information, which can be useful for analysis, prediction, recommendations, and decision making. Data mining and machine learning techniques are used to transform the available data into valuable information. In medical science, disease prediction at the right time is the central problem for professionals for prevention and effective treatment plan. Sometimes, in the absence of accuracy this may lead to death. In this study, we investigate supervised machine learning algorithms—Naive Bayes, random forest, and decision tree algorithm—for prediction of anemia using CBC (complete blood count) data collected from pathology centers. The results show that Naive Bayes technique outperforms in terms of accuracy as compared to C4.5 and random forest.

Keywords Anemia · Classification algorithms · Decision making
Complete blood count (CBC)

1 Introduction

The modern health care system generates huge volume of data every day. There is a need to mine and analyze these data to extract useful information and to reveal hidden pattern. Data mining is the process of discovering new patterns from data collected from varying sources. A number of machine learning algorithms have been used successfully in making prediction in various domains such as healthcare, weather

M. Jaiswal (✉) · A. Srivastava · T. J. Siddiqui
Department of Electronics and Communication, University of Allahabad, Allahabad, India
e-mail: manish.jk50@gmail.com

A. Srivastava
e-mail: animasparklestar@gmail.com

T. J. Siddiqui
e-mail: siddiqui.tanveer@gmail.com

A. Khare et al. (eds.), *Recent Trends in Communication, Computing, and Electronics*, Lecture Notes in Electrical Engineering 524,
https://doi.org/10.1007/978-981-13-2685-1_44

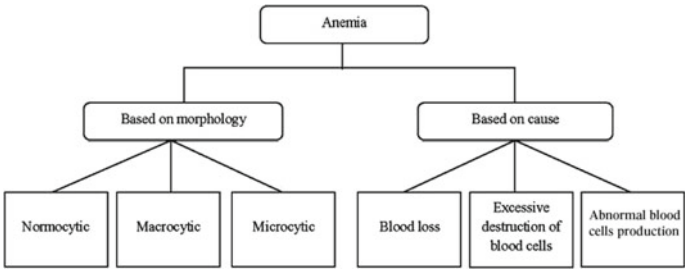

Fig. 1 Classification of anemia

forecasting, stock price prediction, product recommendation. An important aspect of medical science research is the prediction of various diseases and factors that cause them. In medical domain, healthcare data are being used to predict epidemics, to detect disease, to improve quality of life and avoid early deaths [1]. In this work, we investigate three different classification algorithms for its prediction.

Anemia is defined as the decrease in amount of red blood cells (RBCs) or hemoglobin in the blood [2] that has significant adverse health consequences, as well as adverse impacts on economic and social development. Although the most reliable indicator of anemia is blood hemoglobin concentration, there are a number of factors that can cause anemia such as iron deficiency, chronic infections such as HIV, malaria, and tuberculosis, vitamin deficiencies, e.g., vitamins B12 and A, cancer, and acquired disorders that affect red blood cell production and hemoglobin synthesis.

Anemia causes fatigue and low productivity [3–5] and, when it occurs in pregnancy, may be associated with increased risk of maternal and perinatal mortality [6, 7]. According to World Health Organization (WHO), maternal and neonatal mortality were responsible for 3.0 million deaths in 2013 in developing countries.

Anemia disease prediction plays a most important role in order to detect other associated diseases. Anemia disease is classified on the basis of morphology or on the basis of its underlying cause (Fig. 1).

Based on the morphology, anemia is divided into three types, which are normocytic, microcytic and macrocytic. Based on cause, anemia is classified into three types, namely blood loss, inadequate production of normal blood and excessive destruction of blood cells.

In this paper, we attempt to investigate the performance of Naive Bayes, random forest and decision tree algorithm for anemia disease prediction on dataset collected from local pathology centers. The need of this investigation arises from the fact that the underlying cause of the disease varies from one region to another. Although random forest classifier has been earlier investigated for predicting heart and chronic kidney disease, to the best of our knowledge it has not been investigated for anemia disease prediction. This adds novelty to the work.

The rest of the paper is organized as follows:

Section 2 introduces and briefly reviews existing related work. In Sect. 3, we discuss various types of anemia diagnosis tests. Section 4 presents proposed methodology. Section 5 presents experimental details and discussion. Finally, we conclude in Sect. 6.

2 Related Works

In the last decade, numerous data mining and machine learning techniques have been used for anemia disease. Most noted ones are the following:

In [8], SMO support vector machine and C4.5 decision tree algorithm have been used for the prediction of anemia and the performance comparison of the two algorithms is done.

In [9], WEKA is used to get a suitable classifier for developing a mobile app, which can predict and diagnose hematological data comments. The authors compared neural network classification algorithms with J48 and Naive Bayes classifier. The results show that J48 classifier exhibits maximum accuracy.

Dogan and Turkoglu [10] developed a decision support system for detecting iron deficiency anemia using the decision tree algorithm. The algorithm uses three hematology parameters, serum iron, serum iron-binding capacity and ferritin. The evaluation is done on data of 96 patients, and the results were successfully matched with physician's decision.

Abdullah and Al-Asmari [11] experimented with WEKA algorithms: Naive Bayes, multilayer perception, J48 and SMO in an attempt to predict anemia types using CBC reports. The evaluation was done on real data constructed from CBC reports of 41 anemic persons. Similar to [9], J48 decision tree algorithm along with SMO was the best performer with an accuracy of 93.75%.

Unlike the work in [9, 11], we have chosen a different set of classifier and local data in our work.

3 Diagnostic Tests Classification

There are four main tests that are ordered to diagnose anemia disorder which are complete blood count (CBC), ferritin, PCR (polymerase chain reaction) and hemoglobin electrophoresis.

- CBC test is the most frequently blood test to measure overall health and determine a wide range of diseases [8] including anemia, infection and leukemia. A complete blood count test measures almost 15 parameters including: hemoglobin (Hb), red blood cells (RBC), hematocrit (HCT), mean corpuscular hemoglobin (MCH), mean corpuscular volume (MCV), and so on [8].

- A ferritin test measures the amount of iron store in the body. High levels of ferritin indicate an iron storage disorder, such as hemochromatosis. Low levels of ferritin indicate iron deficiency, which causes anemia.
- PCR test is a molecular test, which is used to diagnose genetic disorder.
- A hemoglobin electrophoresis test is a blood test used to measure and identify the different types of hemoglobin in the bloodstream.

4 Methodology

We have used three classifiers, namely random forest, Naive Bayes and decision tree C4.5 algorithm. Figure 2 depicts the flowchart of the proposed method.

4.1 Random Forest Algorithm

Random forest (RF) algorithm derives from decision tree classifier. It is a combination of tree predictors, which aggregates the results of all the trees in the collection and uses majority voting in prediction.

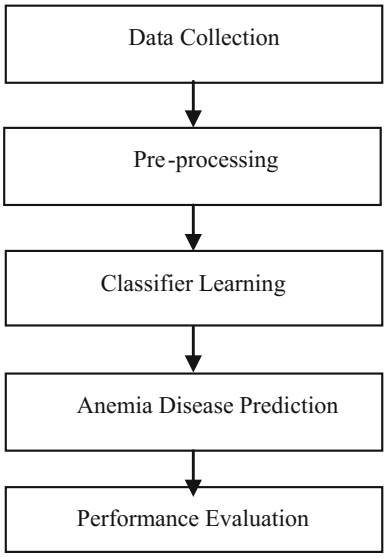

Fig. 2 Flowchart of proposed model

4.2 Decision Tree Algorithm

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. It has been extensively used in various fields [12, 13]. C4.5 (J48 in WEKA) is a decision tree developed by Ross Quinlan.

4.3 Naive Bayes Algorithm

Naive Bayes algorithm is based on Bayes rule of conditional probability. It uses all the attributes contained in the data and analyzes them individually as they are equally important and independent of each other. It requires very less amount of training data.

5 Experimental Results and Discussion

5.1 Dataset

We collect data from different pathology centers and laboratory test centers in nearby area. The collected dataset consists of 200 test samples. These are CBC test data. The dataset contains 18 attributes out of which we have selected only those that are required for anemia disease detection. These are age, gender, MCV, HCT, HGB, MCHC and RDW.

5.2 Experimental Setup

The proposed method uses CBC test values. First, the data are pre-processed to extract the seven attributes as mentioned in Sect. 5.1. Then, we apply the random forest, decision tree and NB classifier on it. The performance evaluation is done in terms of accuracy and mean absolute error (MAE). The mean absolute error (MAE) measures how close the predictions are to the eventual outcomes. Table 1 shows the results of the three classifiers. Tenfold cross-validation has been used to obtain accuracy.

The comparative performance of each classification algorithm based on accuracy and MAE is shown in Figs. 3 and 4, respectively. The Naive Bayes classifier exhibits the best performance on our dataset, which is unlike [9] and [11]. It is not surprising because the dataset being used in these works is different and the cause of disease in different countries might be different. We achieve a maximum accuracy of 96.09%

Table 1 Comparision of algorithms

	Random forest	Naive Bayes	C4.5
Mean absolute error	0.0332	0.0333	0.0347
Accuracy	95.3241	96.0909	95.4602

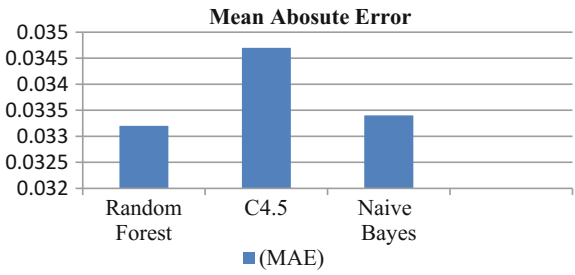

Fig. 3 MAE using each algorithm

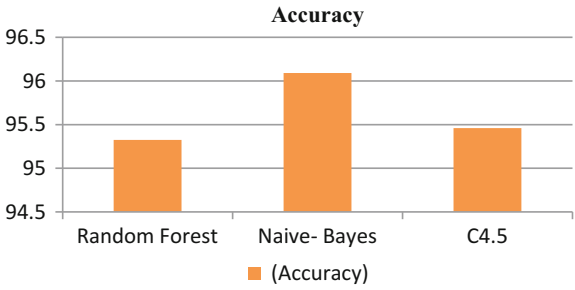

Fig. 4 Comparison of accuracy using each algorithm

with NB classifier which is better than the best performing classifiers—SMO and J48 with an accuracy of 93.75%—reported in [11].

6 Conclusion and Future Work

In this paper, we have compared the performance of three different classifiers in the prediction of anemia disease. The experimental result on a sample dataset suggests that Naive Bayes classification algorithm provides the best performance in terms of accuracy as compared to C4.5 and random forest. Automatic prediction can reduce manual effort involved in diagnosis. In the future, automated tools can be developed which can help the prediction results to suggest further diagnosis. Such automated tools can prove valuable in timely detection of more serious diseases. Furthermore, such disease prediction system can be extended to recommend a treatment plan.

References

1. Arun, V., et al. (2015). Privacy of health information in telemedicine on private cloud. *International Journal of Family Medicine & Medical Science Research*.
2. Provenzano, R., Lerma, E. V., & Szczech, L. (2018). *Management of anemia*. Springer.
3. Ezzati, M., Lopez, Ad, Rodgers, A., & Murray, C. J. L. (2004). *Comparative quantification of health risks: Global and regional burden of disease attributable to selected major risk factors*. Geneva: World Health Organization.
4. Balarajan, Y., et al. (2011). *Anaemia in low-income and middle-income countries*.
5. Haas, J. D., Brownlie, T. (2001). Iron deficiency and reduced work capacity: A critical review of the research to determine a causal relationship. *The Journal of Nutrition*.
6. Kozuki, N., Lee, A. C., & Katz, J. (2012). Child health epidemiology reference group. Moderate to severe, but not mild, maternal anemia is associated with increased risk of small-for-gestational-age outcomes. *The Journal of Nutrition*.
7. Steer, P. J. (2000). Maternal hemoglobin concentration and birth weight. *The American Journal of Clinical Nutrition*.
8. 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.
9. Amin, N., & Habib, A. (2015). Comparison of different classification techniques using WEKA for hematological data. *American Journal of Engineering Research, 4*(3), 55–61.
10. Dogan, S., & Turkoglu, I. (2008). Iron deficiency anemia detection from hematology parameters by using decision tree. *International Journal of Science and Technology*, 85–92.
11. Abdullah, M., & Al-Asmari, S. (2016). Anemia types prediction based on data mining classification algorithms. In *Communication, management and information technology*. London: Taylor & Francis Group.
12. Jerez-Aragonés, J. M., et al. (2003). A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artificial Intelligence in Medicine*, 45–63.
13. Podgorelec, V., et al. (2002). Decision trees: An overview and their use in medicine. *Journal of Medical Systems*, 445–463.