Prediction of Anemia Disease Using Machine Learning Algorithms

Aditya Dixit, Rahul Jha, Raunak Mishra, and Sangeeta Vhatkar

Abstract As we know, Red Blood Cells are the main part of blood that is responsible for the circulation of blood in the human body. Anemia is a well-known disease that is caused due to the deficiency of healthy red blood cells. Due to anemia, red blood cells are unable to supply oxygen throughout the body. This sickness can be lethal to the human body if not treated promptly. We are using machine learning techniques such as Random Forest, SVM, and others to detect anemia in a patient in this study. We can detect anemia in a patient using machine learning methods. As a result, we intend to create a classification-based ML model in which we provide the essential CBC test values for our model to predict whether a patient is anemic. With the help of machine learning techniques, we are automating the process for detecting anemia in this study work. We compared the statistical analysis of all algorithms we've utilized to predict anemia in this paper.

Keywords Anemia · Machine learning · Random forest · SVM · Naive Bayes

1 Introduction

There has been an exponential increase in the data generated through the health industry because of the remarkable advances in Technology used. Using this data, we can extract all the useful information which can then be used for analysis, recommendation, prediction and decision making. In medical science, disease prediction at the right time is important for prevention and effective treatment plan. Anemia

A. Dixit (✉) · R. Jha · R. Mishra · S. Vhatkar
IT Department, Thakur College of Engineering and Technology, Mumbai, India
e-mail: 1032190284@tcetmumbai.in

R. Jha
e-mail: 1032190307@tcetmumbai.in

R. Mishra
e-mail: 1032190325@tcetmumbai.in

S. Vhatkar
e-mail: sangeeta.vhatkar@thakureducation.org

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

V. E. Balas et al. (eds.), *Intelligent Computing and Networking*, Lecture Notes in Networks and Systems 632, https://doi.org/10.1007/978-981-99-0071-8_18

is a disease which is caused by the deficiency of healthy red blood cells which are unable to deliver oxygen throughout the body. Anemia is highly prevalent in India. The third National Family Health Study (NFHS-3) conducted during 2005–6 found that amongst children aged 6–59 months, the prevalence of anemia is 69.5%; in rural India, the prevalence is 71.5%. The prevalence of anemia is maximum among younger children between the age of 12–17 months and 18–23 months. The prevalence of anemia in rural areas appeared to have risen since the previous NFHS (in 1998–9) [1].

Hence, it is important to take some measures to prevent the spread of anemia as much as possible using the latest advancements happening in the Tech Industry. In our study, we found out using various classifier algorithms like Random Forest, SVM, Naïve Bayes etc., we can predict the early stage of anemia so that patients can take required medicine on time and prevent anemia [2]. This project is important as, using the latest advancement in the field of machine learning, we can also make solutions in the field of medical science. This technology can be used in many areas like rural areas where health care systems are still not developed to the extent that of urban areas [3, 4].

Anemia is a disease, which needs timely treatment and early diagnosis, using machine learning we can achieve this. Machine Learning can help us overcome many different problems faced by our country in the field of medicine. Using this project, we will be able to detect whether a person or patient is suffering from anemia or not in a matter of seconds [3].

2 Problems Faced

Anemia is a growing problem amongst young children living in rural India. In Rural areas, there is a lack of proper medical treatment and experienced doctors. This leads to patients traveling long distances to visit experienced doctors for treatment. This delay ultimately leads to the disease becoming more fatal.

Also, many people avoid going to the doctor because they are scared or they can't afford it. Also, due to the lack of trained or experienced doctors in rural areas, they misdiagnose the symptoms resulting in Anemia becoming more fatal.

Anemia, also goes quite unnoticed in many people especially children, which can go unnoticed at first but suddenly become fatal in nature. To identify this, a doctor needs to go through the CBC blood test report thoroughly to identify the early stages of Anemia. Once identified, it is quite easy to cure the disease.

To tackle all these problems, we are planning to create a Machine Learning Model, using which we would make use of multiple algorithms like Naive Bayes, Random Forest, SVM, etc. and select the best algorithm using which we will create a website, where the user can simple put in their blood test parameters in our machine learning model which would then predict whether the user is suffering from Anemia or not.

Our machine learning model can predict and alert the user if the user is suffering from anemia and using which the user can be treated on time without the need of any experienced medical staff.

3 Methodology

We followed the below methodologies to make our project:

1. Taking Input Data
   - Firstly we collect the dataset [5].
   - Dataset should be in csv format (Fig. 1).
   - We import the dataset using various python libraries like Pandas.
   - Above, in our dataset, we have considered five parameters—[3, 6, 7].
     1. Gender—Gender is a very important parameter as the blood parameters and limits for both Male and Female are different and vary, so it is important to also consider this factor.
     2. MCV—MCV stands for mean corpuscular volume. Basically this blood test measures the average size of the red blood cells. Using this test we can get to know whether our red blood cells are too small or too large which can depict any blood disorder such as anemia [8].
     3. MCH—MCH is short for "mean corpuscular hemoglobin." It's the average amount in each of your red blood cells of a protein called hemoglobin, which carries oxygen around your body [9].
     4. MCHC—MCHC is a similar measure to MCH, MCHC stands for "mean corpuscular hemoglobin concentration". MCHC checks the average amount of hemoglobin in a group of red blood cells. A doctor might use both MCHC and MCH in order to diagnose Anemia [10].
     5. Hemoglobin—This parameter tells us about the amount of oxygen present in our blood. It is basically a protein which has the capacity to carry oxygen throughout the body from the lungs. It is also a very important parameter in prediction of anemia. For men, anemia is typically defined as a hemoglobin level of less than 13.5 g/dl and in women as hemoglobin of less than 12.0 g/dl [11].
2. Pre-processing and Cleaning Dataset

	Gender	Hemoglobin	MCH	MCHC	MCV	Result
0	Male	14.9	22.7	29.1	83.7	Not Anemia
1	Female	15.9	25.4	28.3	72.0	Not Anemia
2	Female	9.0	21.5	29.6	71.2	Anemia
3	Female	14.9	16.0	31.4	87.5	Not Anemia
4	Male	14.7	22.0	28.2	99.5	Not Anemia
5	Female	11.6	22.3	30.9	74.5	Anemia

Fig. 1 Picture of anemia CBC dataset [5]

- For data cleaning and preprocessing, we have imported the required dataset using the pandas dataset.
- After importing the dataset and making it a dataframe, we have first converted all values into integers. Checked for null values, we didn't find any null values in our dataset.
- Next, we went ahead and checked all the number of entries and removed all duplicates.
- Now, after cleaning the data, we went ahead for data visualization (Figs. 2 and 3).

3. Feature Extraction/Feature Selection

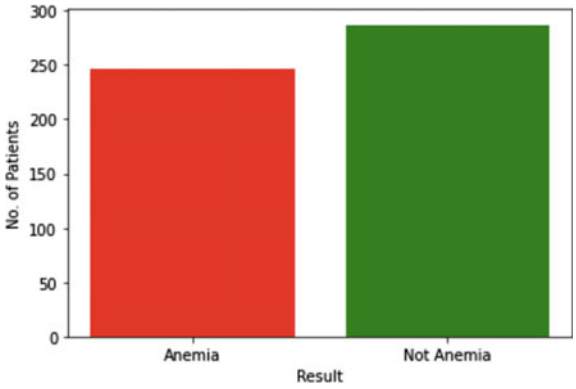

Fig. 2 Split of results in dataset [5]

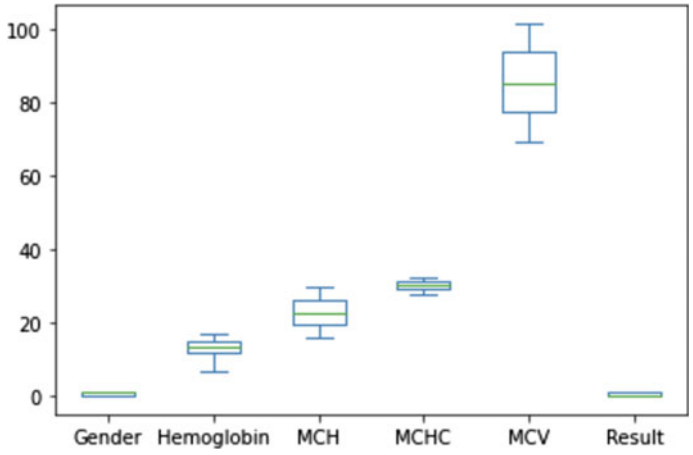

Fig. 3 Boxplot of all the parameters [5]

- As discussed above, we are using 5 features to predict whether a user/patient is suffering from anemia or not.
- We are using Gender, Hemoglobin, MCH, MCHC and MCV from the blood test reports to predict whether a user is suffering from anemia or not [3].
- After cleaning all the data, we will then Normalize the data using MinMaxScaler. MinMaxScaler transforms all the features between 0 and 1.

Here we extracted features that are required for model training (Fig. 4).

4. Apply Classification Algorithms

- After feature extraction now comes to model training.
- First of all we have divided the dataset into training and testing using a method called train_test_split(). We have divided our dataset into a 75–25% train-test split.
- Now select the classification algorithm and import it from respective libraries.

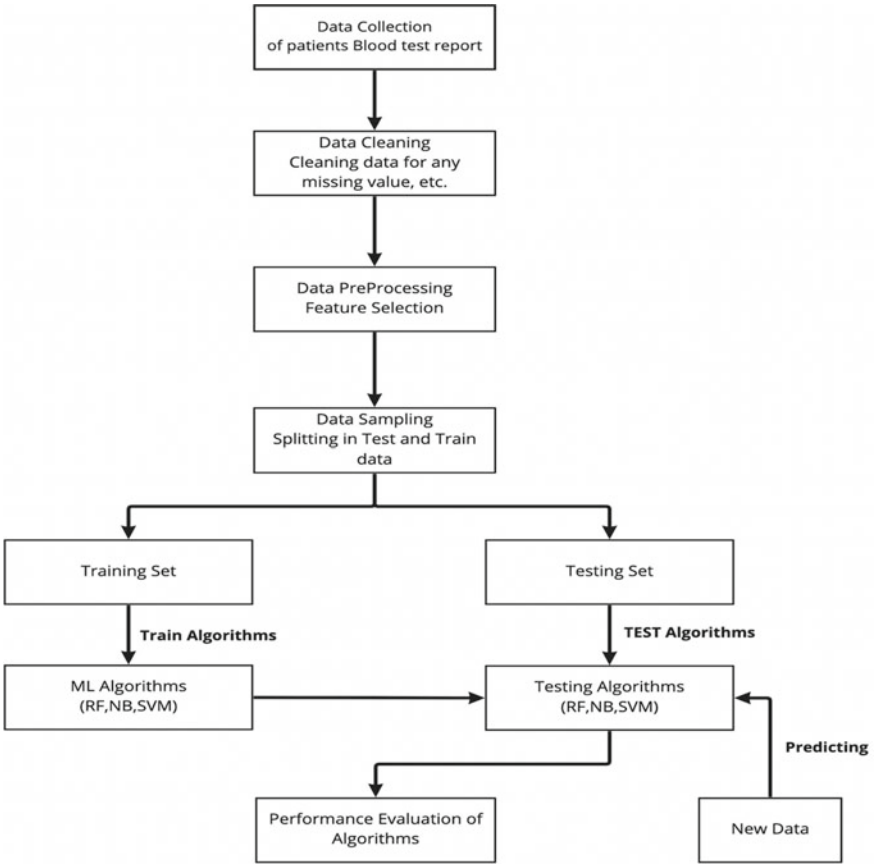

Fig. 4 Flowchart [2, 4]

- Algorithms that we are going to use are Random Forest, SVM, Naïve Bayes etc. [2, 3].
- Below are the detailed study of our algorithms
  1. *Naive Bayes Algorithm*—Naive Bayes Algorithm is a supervised machine learning algorithm which is based on the famous bayes theorem and is used mostly to solve classification problems. It is one of the easiest and effective classification algorithms. It basically predicts the output based on the basis of the probability of the object [12].

     Now, defining the formula as per our project

     P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

     P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

     P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

     P(B) is Marginal Probability: Probability of Evidence [12].

$$\mathbf{P(A|B)} = \mathbf{P(B|A)\,P(A)/P(B)} \tag{1}$$

     As per our problem, We define the formula:

     P(YES|Anemia) is Probability of having Anemia Disease in a person.

     P(Anemia|YES) is the value of patients having parameters outside the normal range having anemia.

     P(Anemia) is the value of people having Anemia.

     P(YES) is the value of total people having blood parameters out of range.

     So, we can rewrite the Naive Bayes algorithm as

     P(YES|Anemia) = P(Anemia|YES) * P(Anemia)/P(YES)

     We will then compare this value with the normal or parameter of people not having Anemia

     P(NO|Anemia) = P(Anemia|NO) * P(Anemia)/P(NO)

     After this calculation, we will in the end compare both these, and the one grater will be the final answer

     If P(NO|Anemia) is greater than P(YES|Anemia), then the person is not suffering from Anemia, else vice-versa.
  2. *Random Forest*—Random forest is a simple to use machine learning algorithm that delivers a good result much of the time, it also does not require us to use hyper parameter tuning. It is also one of the most commonly used algorithms due to its simplicity and versatility which can be used as both regression and classification algorithms [13].

     Why use Random Forest?

     Random Forest is one of the most popular machine learning algorithms used for both classification and Regression problems. It is used because of its speed, that is it works very fast even for very big datasets. It

also provides a very high accuracy in comparison with the other machine learning algorithms [13].

How does the Random Forest algorithm work?

Random Forest as the name suggests, is an algorithm created by the use of multiple decision trees. In this, Random Forest algorithm creates multiple decision trees, and then as per the input, the decision tree shows the output. In random forest, the algorithm for a classification problem takes all the majority classes predicted by all decision trees and average of all predicted outputs for a Regression Problem [13].

Now, lets see the working of Random Forest Algorithm

Step-1: Firstly, we select random data points from the training data set.

Step-2: Next, we build a decision tree for each of the respective data points.

Step-3: Next, we decide the number of decision trees we want.

Step-4: Repeat Step 1 and 2.

Step-5: Now, for predicting, compile all the outputs of all decision trees and take the majority of all outputs for the final output.

3. *SVM*—Support Vector Machine is one of the best machine learning algorithms when it comes to classification problems. This is exactly what SVM does! It tries to find a line/hyperplane (in multidimensional space) that separates the two classes. It then classifies the new point as to whether it lies on the positive or negative side of the hyperplane, depending on the classes to be predicted [14].

   Steps to implement support vector regression in python:

   - Import the library
   - Read the dataset
   - Feature Scaling
   - Fitting SVR to the dataset
   - Predicting a new result
   - Visualizing the results of support vector regression

   Support vector regression is the counterpart of a support vector machine for regression problems. In our project we are using different attributes of the dataset and predicting the result using this support vector machine [14].

5. Real Time Implementation of Project

   - Here comes the main part where we have to map our project with the real world problems.
   - For this purpose we are trying to reach the various resource persons which are pathologists/doctors and provide them with solutions that our model is giving.
   - We have decided to provide our service to NGOs or Social work bodies or organizations or medical bodies or rural clinics or hospitals where there is a lack of experienced medical staff.

- Patients can, on our website, just put in their blood test parameters and our machine learning model will predict whether the patient is suffering from anemia or not.

4 Technology Used

Technology and Tools that we are going to use in out project:

We are using one of the most useful and powerful language i.e. python. Python also has robust library support for Machine learning.

1. Google Collab—This is a jupyter notebook IDE where we can easily run and also see the output of each cell simultaneously. We will use Google Colab as it already has many of the required libraries installed.
2. Pandas—This is one of the most important libraries for data science applications. It is used for cleaning and perfecting our dataset before inserting it in the machine learning model.
3. Scikit—It is a machine learning library containing many models like classification, regression and clustering algorithms. It also has a metrics module which is used for checking the accuracy of the models.
4. Matplotlib—It is a library used for data analysis. It is a library used to create various types of graphs.
5. Seaborn—It is a library used for creating many types of graphs.
6. Flask—It is a library which we will use to create our website where the user enters their CBC parameters.

5 Result and Discussion

- After implementation of all the above steps, we have come up with the accuracy we have achieved using the Random Forest, Naive Bayes and SVM algorithms (Table 1).
- Above is the accuracy we have achieved from our algorithms after training them and then testing them with the test data.
- We have also below attached the True Positive, True Negative, False Positive and False Negative of each algorithm.

Table 1 Algorithm accuracy

Algorithm	Accuracy (%)
Random forest	99.38
Naive Bayes	95.65
SVM	97.52

Algorithms	True positive	True negative	False positive	False negative
Random forest	99	61	1	0
Naive Bayes	95	59	5	2
SVM	97	60	3	1

- Using TP, TN, FP, FN we have found the accuracy using the formula [15]

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

- As we see, our results are up to standards and the accuracy of each algorithm is very good, even exceeding our expectations.

Acknowledgements The success and final outcome of this project required a lot of guidance and assistance from many people and we are extremely privileged to have completed the project successfully. We would like to thank everyone for their guidance.

We sincerely thank our Principal, Dr. B. K. Mishra, Vice Principal, Dr. Kamal Shah and HOD IT, Dr. Sangeeta Vhatkar for always encouraging us to do our best. We are highly indebted to our guide Dr. Sangeeta Vhatkar who supported and constantly supervised us through this project and helped us in not only completing this project but also provided us with a sample amount of knowledge that was really beneficial to us.

We are thankful to and fortunate enough to get constant encouragement, support and guidance from all teaching staff of the IT Department who helped us in successfully completing our project work. Also, we would like to extend our sincere thanks to all staff in the laboratory for their timely support.

We would like to express our gratitude towards our parents for their kind cooperation and encouragement which helped us in completion of this project.

References

1. Pasricha SR, Biggs BA, Prashanth NS, Sudarshan H, Moodie R, Black J, Shet A (2011) Factors influencing receipt of iron supplementation by young children and their mothers in rural India: local and national cross-sectional studies. BMC Public Health 3(11):617. https://doi.org/10.1186/1471-2458-11-617.PMID:21810279;PMCID:PMC3171369
2. Jaiswal M, Siddiqui TJ (2018) Machine learning algorithms for anemia disease prediction: select proceeding of IC3E 2018. https://www.researchgate.net/publication/329484705
3. Pavan B, Chandra YH, Yeruva S, Shradhah M, Jain S, Kumar AR, Kondaveeti S (2020) Prediction of anemia disease using classification methods. EasyChair Preprint April 13, 2020
4. Yıldız TK, Yurtay N, Öneç B (2021) Classifying anemia types using artificial learning methods. Eng Sci Technol Int J 24(1):50–70. ISSN 2215-0986. https://doi.org/10.1016/j.jestch.2020.12.003. (https://www.sciencedirect.com/science/article/pii/S2215098620342646)
5. https://www.kaggle.com/code/rahulsarkar221/anemia-predictive-analysis/data
6. Barpanda SS (2013) Use of image processing techniques to automatically diagnose sickle-cell anemia present in red blood cells smear. Department of Electrical Engineering National Institute of Technology Rourkela-769008 (ODISHA), May-2013. https://core.ac.uk/reader/5318955
7. Abdullah M, Al-Asmari S (2016) Anemia types prediction based on data mining classification algorithms. Communication, management and information technology. CRC Press, pp 629–636. https://www.researchget.net/publication/311107778,2016

8. https://medlineplus.gov/lab-tests/mcv-mean-corpuscular-volume/
9. https://www.webmd.com/a-to-z-guides/what-are-mch-levels
10. https://www.healthline.com/health/low-mchc
11. https://medlineplus.gov/lab-tests/hemoglobin-test/
12. https://www.javatpoint.com/machine-learning-naive-bayes-classifier
13. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
14. https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm
15. https://www.javatpoint.com/confusion-matrix-in-machine-learning