A Smart Way to Detect Health Conditions from Blood Report using Machine Learning

[1]**N. Ramesh Babu**
rameshbabun@rguktsklm.ac.in

[2]**K. Manoj Kumar**
s170936@rguktsklm.ac.in

[3]**B. Prem Kumar**
s170456@rguktsklm.ac.in

[4]**N. Kishorekumar**
s170095@rguktsklm.ac.in

[5]**K. Umamaheswari**
s170314@rguktsklm.ac.in

Date: March 20, 2023

Department of ***Electronics and Communication Engineering***

Rajiv Gandhi university of Knowledge Technologies,
Srikakulam

Date: March 20, 2023

Contents

Abstract

This project aim is to detect complications in the blood. Nowadays blood analysis plays a major role in the pharmaceutical world. Blood analysis is very essential to treat any disease. Blood report consists of various measures like red and white blood cells, platelets count and hemoglobin etc. The blood analysis report is one of the main reports in the medical field to analyse the health of a patient. Doctors use this report to understand the health condition of a patient to give medication and to perform any surgery. In the Blood measure report, every measure has some range to maintain our body healthy. By this range, we can get an idea about which measure is not in range or deficit in range. So, it is a very crucial thing to understand the condition accurately and effectively. Nowadays, most people are becoming sick and hospitals are becoming crowded. At that time the hospital staff or doctors can upload the report to get a perfect analysis. The project aims to extract the information from the uploaded report using OCR and to apply an ML algorithm to that extracted values to detect complications in the blood and gives the result through the designed website. [1]

Keywords:

Blood report, Optical Character Recognition (OCR), ML algorithm, Web-Framework, file operations

I INTRODUCTION

Blood is one of the major components of the human body. Like organs, blood is very important. Blood contains different types of parameters like Haemoglobin, Red blood cells, White blood cells, Platelets etc.., Each parameter has its range of values. The changes in these values can affect the human and may cause diseases. Changes in different parameters can cause different types of diseases. There are various types of diseases, which can be caused by changes in the blood. The range of values will vary based on age and gender also.

Analysing the blood report is the major step to detect the health condition of a patient. Most of the blood tests don't need special conditions like fasting etc., Blood tests can be done at any time. Different parameters can be measured by testing blood. The results will help to identify health problems in the early stages. Only a blood test is not enough to deal with a patient's health condition, but it's one of the factors. Applying modern technology to our daily life work will reduce human effort and can increase accuracy. Applying modern technologies like Machine Learning and Artificial Intelligence is a hot topic nowadays

Machine Learning is a data analysis technology that teaches computers to act like humans. It's a data-driven technology, it works based on the dataset given to it. The performance of the machine learning algorithm can be improved according to the dataset provided. The main objective of this project is to predict the health condition of a patient from their blood report. Different Machine Learning algorithms are used to get an algorithm with maximum accuracy.

Getting results through an interface is also important for the project. Creating 2 webpages, in which, one is to upload our document or photo of the report and another is to show the results.

The rest of the paper is organized as follows. Section II introduces background information about the used techniques. Section III presents the different related methods of blood disease prediction using ML classifiers. Section IV describes the data set and the blood test attributes. Section V shows the results of the experiment. Section VI presents the website, which we are going to display results through. Finally, section VII presents the conclusion and future work of the research.

II BACKGROUND

Machine learning is a computer science branch that is responsible for the development of computer systems that can learn and change their actions according to the situation. Machine Learning technology is depending on learning from the data-set given to it and evaluating the model results and trying to optimize the output. Fig.II Shows a brief detailing of the machine learning activity. There are three types of algorithms. They are:

- Supervised Learning: The computer trained with presented inputs and their desired outputs, for predicting the output of future inputs.
- Unsupervised Learning: The computer presented with inputs without desired outputs.
- Reinforcement learning: The computer interacts with the environment, and it must perform a specific goal without training.

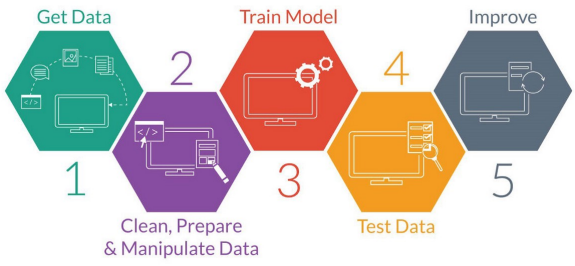

Figure 1: Machine Learning Activity

Machine Learning techniques become an essential tool for prediction and decision-making in many disciplines. The availability of clinical data leads machine learning to play a critical role in medical decision-making. It serves as a valuable aid in identifying a disease for improving clinical decisions and choosing suitable medical procedures.

III CLASSIFIERS USED

We used the following classifiers for classifying the patients based on learning datasets; these classifiers are:

- Random forests classifier: it is a band learning method for classification that operates by constructing a multitude of decision trees by training records with their labelled classes. After building the tree, the unknown records could be classified.

- Support vector machine: it represents the training data as points in a flat separated space by an apparent gap. New examples are mapped into space with the forecast category based on which side of the gap they fall.

- Decision Tree: it models the attributes and their values with decisions in the tree; where the nodes contain attributes with their values and leaves contain decisions. The algorithm considers all features and makes a binary split on them. It orders the attributes on the tree according to the information gain value in descending order. After building the tree, new tuples will be classified according to their values by traversing the tree until reaching the leaf that contains the class.

All these classifiers are used in the disease prediction process for improving clinical decision-making and minimizing medical errors, in the next section, we listed the recent research that uses machine learning in blood disease analysis.

IV DATA SET and DISEASES

The dataset on which the ml model is trained is created after a lot of research. Creating a dataset is the major part of this project. The dataset contains 4,500 data points and 15 classes. Each class contains 300 data points. Each class has values based on the age and gender of the patient. Each class represents each disease. These are the diseases that ml model is going to predict.[2]

- **Anaemia:** it is a decrease in the amount of haemoglobin or red blood cells in the blood. It may cause vague and may include feeling tired, shortness of breath, or weakness.[3] Based on the count, it has four types:

  1. Mild Anaemia
  2. Moderate Anaemia
  3. Severe Anaemia
  4. Dangerous Anaemia

- **Polycythaemia:** an abnormally high amount of haemoglobin or red blood cells in the blood.

- **Thrombocytopenia:**it is about the lack of platelets. It is not so dangerous but sometimes leads to bleeding too much. Based on the count, it has four types:

  1. Mild Anaemia
  2. Moderate Anaemia
  3. Severe Anaemia
  4. Danger

- **Thrombocytosis:** an abnormally high number of Platelets in blood.

- **Leucocytosis:** it causes an increase in white cells above the normal range in the blood. It may cause certain parasitic infections or a tumour, as well as leukaemia.

- **Leukopenia:** it causes a decrease in white cells below the normal range in the blood.

- **Neutrophilia:** is defined as a higher neutrophil count in neutrophil count in the blood than normal. Neutrophilia can bee is seen in infections and inflammation.

- **Neutrophilia:** It can be caused by diseases that damage the bone marrow, infections or certain medications.

- **Eosinophilia:** High eosinophilia levels can indicate a mild condition such as a drug reaction or allergy.

- **Basophilia:** Basophilia may be a sign you have an infection, or it may be a sign of serious medical conditions like leukaemia.

- **Monocytosis:** Having an abnormally high number of infection-fighting monocytes. It may signify a severe medical condition such as an autoimmune disease, a blood disorder, or cancer. It may also mean that you have encountered an infection.[4]

- **Lymphocytosis:** Causes when we have high lymphocytes count. Lymphocytosis is one of the first signs of certain blood cancers.

- **Normal:** in this class, all parameters' values are normal, and there are no essential notifications in the blood analysis.[5]

Parameters	Description
Age	Age of the patient
Gender	Gender of the patient
HGB	Haemoglobin
Thrombocytes	Platelets Count (in Millions)
Leukocytes	White Blood cells count
Neutrophil	Per cent Neutrophils in blood
Eosinophil	Per cent of Eosinophils in blood
Basophil	Per cent of Basophils in blood
Lymphocyte	Per cent of Lymphocytes in blood
Monocyte	Per cent of Monocytes in blood

Table 1: BLOOD ANALYSIS PARAMETERS

V EXPERIMENT RESULTS AND DISCUSSION

Cross-validation is a statistical method of evaluating and comparing learning classifiers by dividing data into parts: one is training data, which is used to train the model. Second is testing data, the model is used to validate the model. The training and testing sets must cross over in successive rounds such that each data point has a chance of being validated.[6]

For each classifier, accuracy is measured. The Random Forest algorithm provides high accuracy and SVM provides less accuracy than Random Forest. The overall results prove the success of applying classical machine learning algorithms in the process of blood disease prediction.

Classifier	Accuracy
SVM	98%
Decision Tree	99%
Random Forest	100%

Table 2: Accuracy Results for each classifier

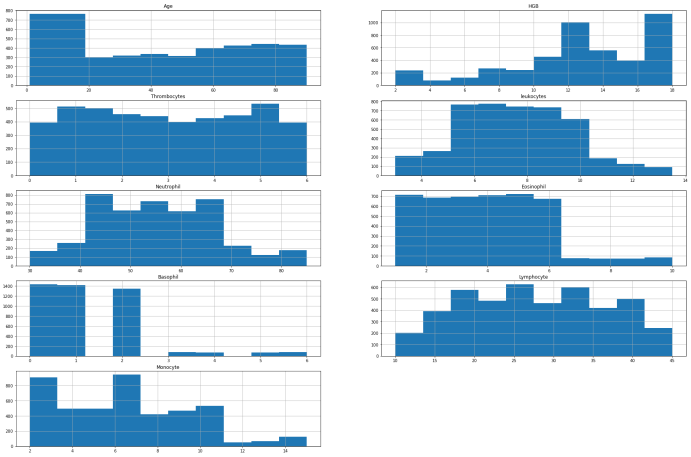

Figure 2: Histogram for each feature

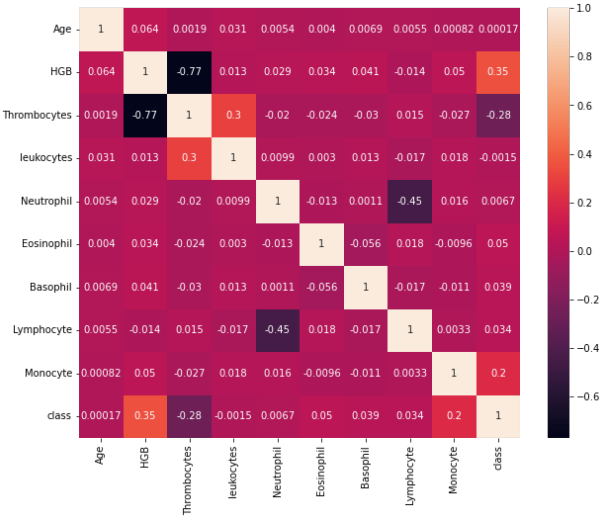

Figure 3: Heat Map

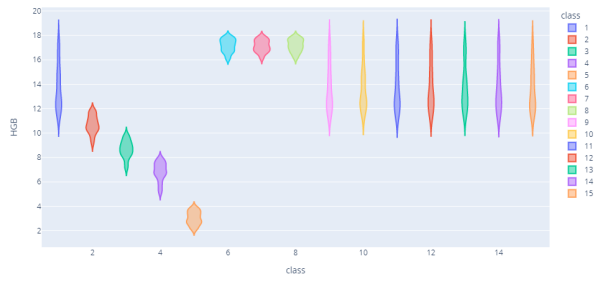

Figure 4: Violin plot for HGB & class

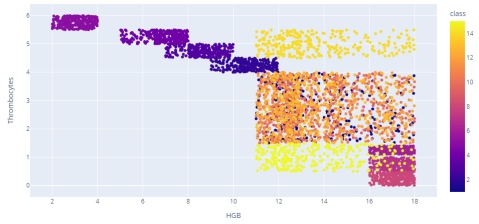

Figure 5: Scatter Plot HGB vs Thrombocytes

VI WEBSITE

We also created a web-page to interact with the patient. A start page, in which we can upload an image or a pdf file. OCR recognizes the desired values from the uploaded files and gives those values to the model. The model predicts the disease using those values and displays the disease along with normal parameters, abnormal parameters, symptoms and precautions through a results page.

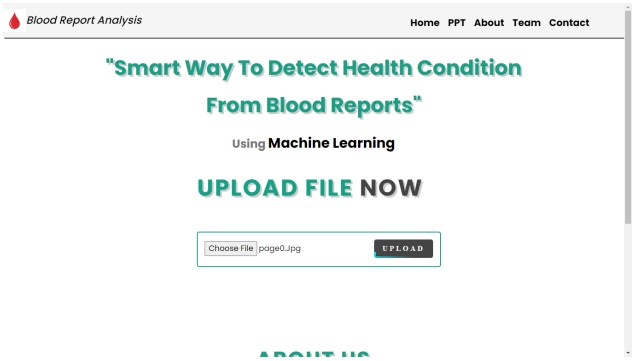

Figure 6: Start Page

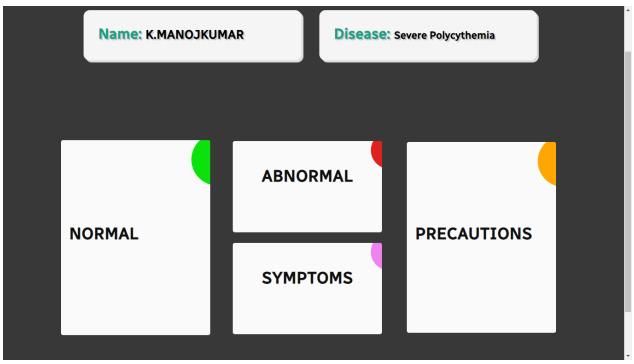

Figure 7: Results Page

VII CONCLUSION AND FUTURE WORK

Machine learning becomes an essential technique for modeling the human process in many disciplines, especially in the medical field, because of the high availability of data. One of the essential disease detectors is the blood analysis; as it contains many parameters with different values that indicate definite proof of the existence of the disease.[6]

The machine learning algorithm accuracy depends mainly on the quality of the dataset; for this reason, a high-quality dataset is collected. This dataset is used for training the classifiers for obtaining high accuracy. We tested several classifiers and achieved accuracy up to 100% which realizes the research objective, which is helping physicians to predict blood diseases according to a general blood test.[7]

The future work will focus on testing the proposed data set using different deep learning algorithms to compare classical and deep learning approaches in this research area.[8]

References

[1] J. E. Allison, I. S. Tekawa, L. J. Ransom, A. L. Adrain, and G. R. Smith, "A comparison of faecal occult blood tests for colorectal-cancer screening," *New England Journal of Medicine*, vol. 334, no. 3, pp. 155–160, 1996.

[2] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[3] F. Cabitza, R. Rasoini, and G. F. Gensini, "Unintended consequences of machine learning in medicine," *JAMA*, vol. 318, no. 6, pp. 517–518, 2017.

[4] A. M. Darcy, A. K. Louie, and L. W. Roberts, "Machine learning and the profession of medicine," *JAMA*, vol. 315, no. 6, pp. 551–552, 2016.

[5] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu, "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 601–606, 2011.

[6] R. S. Michalski and Y. Kodratoff, "Research in machine learning: Recent progress, classification of methods, and future directions," in *Machine Learning*. Morgan Kaufmann, 1990, pp. 3–30.

[7] S. H. Park, C. Park, Y. Kim, J. Y. Han, and C.-J. Park, "Establishment of age-and gender-specific reference ranges for 36 routine and 57 cell population

data items in a new automated blood cell analyzer, sysmex xn-2000," *Annals of Laboratory Medicine*, vol. 36, no. 3, pp. 244–249, 2016.

[8] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.