INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

OPEN ACCESS

Received: 19-10-2024

Accepted: 16-12-2024

Published: 08-01-2025

Citation: Gaikwad DK, Mahale V, Gaikwad A, Gaikwad A (2025) Automatic Blood Disease Identification System Using Machine Learning Algorithms with Hemogram Reports. Indian Journal of Science and Technology 18(1): 11-22. https://doi.org/10.17485/IJST/v18i1.3443

* **Corresponding author**.

deepa.gaikwad76@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Automatic Blood Disease Identification System Using Machine Learning Algorithms with Hemogram Reports

Deepali K Gaikwad[1]*, **Vivek Mahale**[2], **Asha Gaikwad**[3], **Ashok Gaikwad**[4]

1 Research Scholar at Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Chh. Sambhajinagar, 431001
2 Associate Professor, Dept. of MCA, Deogiri College, Chh. Sambhajinagar, 431001, India
3 HOD of Computer Science and Engineering (Data Science), P.E.S . College of Engineering, Chh. Sambhajinagar, 431001, India
4 Director, Institute of Management Studies and Information Technology, Vivekanand College Campus, Chh. Sambhajinagar, 431001, India

Abstract

Objectives: The primary objective of this research is to create prediction models with machine learning algorithms that can detect blood illnesses in their early stages, including sickle cell disease, leukemia, anemia, and lymphoma. **Methods:** Reducing the amount of deaths linked to these illnesses and increasing the precision of disease detection. The study used machine learning algorithms: Decision Tree, Random Forest, and XGBoost classifier, and it used a dataset of complete blood count reports. Based on the characteristics of the total blood count, these algorithms were trained on the dataset to forecast the probability of blood diseases. **Findings:** According to the study, there is a better chance of a cure for blood problems since machine learning models can correctly detect them in their early stages. The findings imply that using predictive models can both lower the death rate and enhance the quality of life for those with these conditions. With a 99.10% accuracy rate, the Random Forest method outperforms other algorithms like Decision Tree, and XGBoost. **Novelty:** This research is unique because it uses a dataset of complete blood count reports results to use machine learning algorithms for blood condition prediction. Healthcare systems may detect blood problems early on and enhance patient outcomes by using classifier algorithms such as Decision Tree, Random Forest, and XGBoost. This opens up new avenues for illness identification.

Keywords: Machine Learning; Decision Tree; Random Forest; XGBoost; Blood Diseases

1 Introduction

The body uses blood as its main fluid for nourishment, oxygen, and waste removal, all of which are critical for human existence[1]. Physicians frequently utilize blood testing as a diagnostic method to assess the health of their patients; the test used will depend on a number of variables, including the patient's gender, symptoms, and underlying medical issues[2]. Blood tests by themselves are unable to make a conclusive diagnosis, but when paired with other clinical indicators, they can provide insightful information that supports a doctor's diagnosis[1]. The medical field of hematology, which focuses on blood abnormalities, includes a variety of illnesses, such as sickle-cell anemia, dengue fever, lymphoma, and anemia.

The importance of machine learning algorithms in medical research, especially in illness prediction, has been emphasized by recent publications[2]. However, the quality of the data collected has a significant impact on how accurate these algorithms are. Prior studies on hematological illnesses have mostly employed conventional machine learning techniques like Decision Tree and Random Forest algorithms to target certain ailments, such as hemolytic anemia and anemia[3,4]. However, to detect a wider variety of blood-related disorders, a more thorough method is required.

By using machine learning algorithms, like XGBoost, to identify different hematological illnesses and blood ailments based on hemogram blood test data, this work seeks to close this research gap. This research also highlights the potential of machine learning technologies to improve illness identification and treatment by offering a comprehensive overview of methods for predicting various blood-related ailments. The ultimate objective of this project is to enhance patient outcomes by using machine learning to create hematological disease diagnostic tools that are more precise and efficient.

The goal of this research is to investigate how machine learning algorithms may be used to diagnose a variety of illnesses, with an emphasis on diabetes, heart disease, Parkinson's disease, and anemia[5]. Significant advancements in this field have been made in recent years, as evidenced by a survey of the literature; this study attempts to address these knowledge gaps. This study specifically seeks to:

1) Examine how well machine learning algorithms diagnose blood diseases and hematological problems.
2) Develop a thorough strategy to recognize a wider variety of blood-related disorders.
3) Offer a comprehensive review of techniques for forecasting various blood-related conditions.
4) Emphasize how machine learning techniques may improve illness diagnosis and care.

In order to improve patient outcomes by creating more precise and efficient diagnostic tools, our project intends to meet these goals and add to the body of knowledge already available in the fields of hematology and machine learning.

Mitushi Soni and Sunita Varma, (2020), diabetes was 77% accurately diagnosed using the PIMA dataset from the UCI source[6]. G Abdurrahman, et. al, (2020), The UCI machine learning repository contains a utilized dataset of 188 Parkinson's disease patients from the Cerrahpa Ya Faculty of Medicine. The model's accuracy score was 85.40%[7] and Nasif Wasek Fahim, et. al., (2020), had a 94.87% accuracy rate in identifying individuals who had Parkinson's disease using XGBClassifier[8].

Boshra Farajollahi et al. (2021), A decision tree was utilized to identify diabetes in 286 women of PIMA Indian ancestry who were at least 21 years old. The precision was 79.87% and Random Forest Algorithm found accuracy of 83%[9]. Mai, C. K., et. al., (2021), 1387 patients with 13 parameters were used to detect Anemia from Thalassemia and Sickle Cell Society using a decision tree. It has a 95% accuracy rate and random forest identified 96% accuracy rate[10]. Using the Random Forest Algorithm, Olta Llahaa Amarildo Ristab (2021) indicated an accuracy of 63.75%[11], V. Jackins et al.'s (2021) prediction of heart disease, diabetes, and cancer showed an accuracy for diabetes data of 74.03%, coronary heart disease data of 83.85%, and cancer data of 92.40%[12]. L. Akter et. al., (2021), concluded that 81% of the participants were suffering from Alzheimer's[13] and In 2021, Nikisha Jadhav and colleagues identified Parkinson's illness with a 92% accuracy rate[14] using the XGBoost classifier.

Parth Verma and Vinay Chopra (2022), Using a C4.5 decision tree, anemia was diagnosed with 95.45% accuracy out of 200 instances containing 18 features. In this study, WEKA tools are employed[15]. Boukhatem, Youssef, H. Y., and Nassif, A. B. (2022) employed Support Vector Machine, Multilayer Perceptron, Random Forest, and Naïve Bayes to diagnose cardiovascular illness; among these, the Support Vector Machine model yielded the best results, with 91.67% accuracy[16]. Kartik Budholiya et al. (2022), The University of California, Irvine's online machine learning dataset was used to detect heart illness early on with 91.8% accuracy repository for data mining[17]. Srichand Doki, et. al. (2022), XGBoost was used to detect cardiac disease. It has an accuracy of 85.96%[18]. Dheiver Francisco Santos (2023) achieved a 93.33% detection accuracy rate for Parkinson's disease[18–21].

Although machine learning is becoming more and more popular for diagnosing diseases, prior research has been constrained by the small size of the datasets employed, which makes it impossible to adequately represent the complete population. Furthermore, the research that have already been done have mostly concentrated on a limited number of machine learning methods, failing to investigate other strategies that can raise efficiency levels. Moreover, a thorough comprehension of the algorithms' performance has been impeded by the absence of comparison with varied datasets, resulting in a notable knowledge gap.

This work intends to overcome these constraints by making use of larger and more varied datasets in order to improve the accuracy and generalizability of the results. This study aims to determine the best approach for disease detection and assess the performance of several algorithms on different datasets by looking into a wider variety of machine learning techniques.

This all-encompassing method will give a detailed grasp of the advantages and disadvantages of every algorithm, which will ultimately guide the creation of machine learning models for illness detection that are more precise and effective.

This study is important because it has the potential to resolve current research gaps and advance the creation of better machine learning models for illness diagnosis. This study intends to offer a more thorough knowledge of the function of machine learning in disease diagnosis by resolving the shortcomings of earlier studies, which should ultimately result in better healthcare outcomes.

2 Blood Related Diseases

Blood diseases and disorders can cause problems for a number of blood components, making it difficult for blood to function as intended. There is a genetic component to many diseases and blood disorders. There are additional causes, such as inadequate dietary intake of certain nutrients, negative medication responses, and other illnesses. Anemia and bleeding disorders similar hemophilia are prevalent blood conditions. A lower than normal concentration of hemoglobin and red blood cells in the blood is the cause of anemia. Hemochromatosis patients have elevated blood iron levels. Leukocytosis is caused by an excess of white blood cells (WBC) in the blood compared to normal levels. Thrombocytopenia may ensue due to the blood's low range of platelet counts.

3 Proposed Method

Applying machine learning techniques to model and forecast illnesses based on blood data, such as Random Forest, Decision Tree, and XGBoost. Figure 1 illustrates methodology for automatically detecting blood disease using machine learning algorithms.

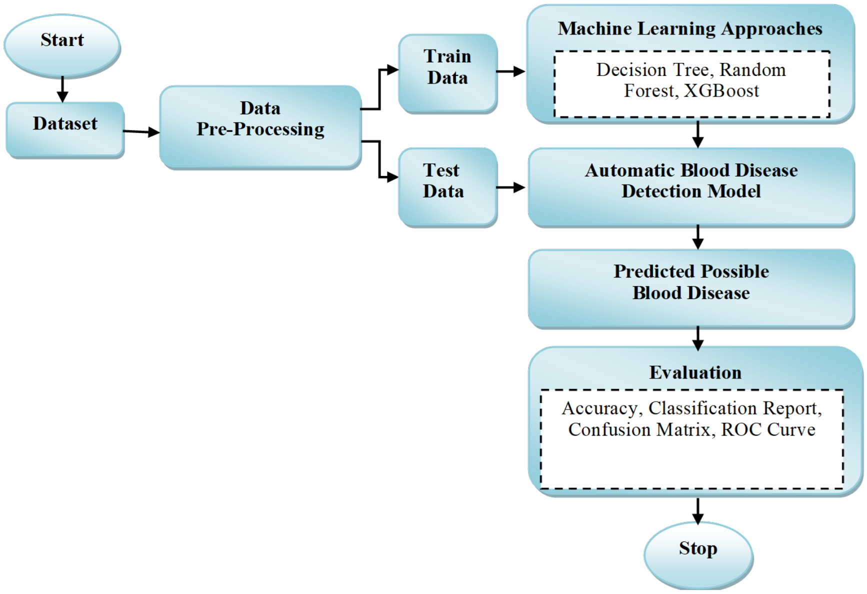

Start
Dataset
Data Pre-Processing
Train Data
Test Data
Machine Learning Approaches
Decision Tree, Random Forest, XGBoost
Automatic Blood Disease Detection Model
Predicted Possible Blood Disease
Evaluation
Accuracy, Classification Report, Confusion Matrix, ROC Curve
Stop

Fig 1. Proposed Method of Automatic Blood Disease Detection System using Machine Learning Algorithms

3.1 Data Set for Blood Diseases Analysis

The complete blood count reports dataset, which included 4451 patients with 18 parameters, was downloaded from kaggle.com. Table 1 lists the eighteen parameters that are present in each blood report[22–25].

3.2 Data Pre- Processing

The data were pre-processed to ensure reliability and remove noise. The Sklearn Python module's simple imputer is used to handle missing values in datasets. The data values were standardized in this step, and any incomplete or erroneous data was

Table 1. Blood Components

Blood Component	Cell	Normal Range as per WHO
White blood cells	WBC	4,000 to 11,000 cells/μL
Neutrophil	NE	2.0 to 7.0 x 10^9 cells per liter (cells/L)
Lymphocyte	LY	1.0 to 4.8 x 10^9 cells/L
Monocyte	MO	0.2 to 1.0 x 10^9 cells/L
Neutrophil Percent	NE %	40% to 75%
Lymphocyte Percent	LY %	20% to 45%
Monocyte Percent	MO %	2% to 10%
Red Blood cells	RBC	Men: 4.5 to 5.5 million cells per microliter (million/μL) Women: 4.0 to 5.0 million/μL
Hemoglobin	HB or HGB	Men: 13.5 to 17.5 grams per deciliter (g/dL) Women: 12.0 to 15.5 g/dL
Hematocrit	HCT	Men: 38.8% to 50.0% Women: 34.9% to 44.5%
Mean Corpuscular Volume	MCV	80 to 96 femtoliters (fL)
Mean Corpuscular Hemoglobin	MCH	27 to 33 picograms (pg)
Mean Corpuscular Hemoglobin Concentration	MCHC	32% to 36%
Red Cell Distribution Width	RDW	11.5% to 14.5%
Platelet	PLT	150,000 to 450,000 platelets/μL
PCT	PCT	0.108 to 0.282 (or 0.108 to 0.282 x $10^6/\mu$L)
MPV	MPV	7.2 to 11.2 femtoliters (fL)
PDW	PDW	9.6% to 15.2%

dealt with and the data values had to be normalized. One popular technique for providing category information in a numerical representation is label encoding. Label encoding is the procedure that converts categorical variables into numerical values. Every single data point in this labeled dataset was linked to the relevant disease class. Training and testing sets were created from the pre-processed data. In this research utilise, 80% of your data for training and 20% for testing. In order to improve the performance of the chosen model, its parameters were calibrated using the training set. This step gave the model knowledge about the relationships between the traits of the blood illness and its effects.

3.3 Machine Learning Approach used by Automatic Blood Disease Detection System

The importance of machine learning approaches for detection and decision making is being recognized by an increasing number of fields. Through the use of machine learning techniques, illness prediction can be finished rapidly and with great accuracy.

3.3.1 Decision Tree

A tree structure is used in this form of classifier, where nodes are features with values, branches are decision rules, and leaves are decisions or outputs. It uses basic decision rules inferred from training data to build a training model that forecasts the value or class of the target characteristics. The decision and the leaf are the two nodes that make up a decision tree. Leaf nodes, or decision nodes, are nodes that make decisions. Decision nodes have many branches, but leaf nodes have no extra branches(2,23,25).

The fundamental step in building a decision tree is choosing the optimal characteristic to divide the data into. This is done by applying metrics such as Gini impurity, entropy, or information gain(26).

Gini Impurity: Indicates the probability that a new instance would be incorrectly classified if it were randomly classified based on the dataset's class distribution.

$$Gini = 1 - \sum_{i=1}^{n} (pi)^2 \tag{1}$$

Where, p_i denotes the likelihood that an instance will belong to a specific class.

Entropy: Quantifies the degree of unpredictability or contamination within the dataset.

$$Entropy = - \sum_{i=1}^{n} pi\, log2(pi) \tag{2}$$

Where, p_i denotes the likelihood that an instance will belong to a specific class

Information Gain: Quantifies the decrease in entropy or Gini impurity following the partitioning of a dataset based on a feature.

$$Information\, Gain = \; Entropy\, parent \;\; \sum_{i=0}^{n} (\left| \frac{Xi}{X} \right| * Entropy(Xi) \tag{3}$$

Where, X_i is X's subset after it has been divided according to a feature.

3.3.2 Random Forest

This study made use of the Random Forest Algorithm. Using training records with their labeled classes, a random forests classifier generates a large number of decision trees. This is a system for grouping band education. After the creation of the tree, the unclassified records may be described[27,28].

The following can be used to summarize the general formula that Random Forest uses to produce predictions:

$$\hat{y} \;=\; mode(\hat{y}1, \hat{y}2, ..., \hat{y}n) \tag{4}$$

Where ŷ is the final prediction and $\hat{y}_i$ are the predictions from individual trees.

3.3.3 XGBoost

”Extreme Gradient Boosting” is what this acronym stands for. Decision trees are used as classifiers in the gradient boosting (GB) method, which is referred to as XGBoost. Scalability, efficiency, and speed are the reasons for its implementation. Here is how GB and XGBoost are often interpreted:

To put it briefly, Gradient Boosting and XGBoost have the following explanations. A dataset of n observations is indicated by the notation D = [x, y], where x stands for the feature (independent variable) and y for the dependent variable Using Ŷi as the forecast for the ith sample at the bth boost, we have the B function to predict the result in GB assuming there is a k amount of boosting. The symbol f_b represents a tree structure q, where leaf j has a weight score w_j.

Finally, by adding together the scores for each leaf, the final prediction for a given sample xi may be found, as illustrated in Equation (1).

$$\hat{y}i \;= \sum_{k=1}^{K} fk(xi) \tag{5}$$

For XGBoost, the goal function L can be expressed as follows:

$$L = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \tag{6}$$

3.4 Automatic Blood Disease Detection System

Created and built Artificial Intelligence Predictive Models [Figure 2] and Algorithms for Blood Disease Detection Using Haemogram Report Datasets. For this objective, several machine learning architectures were taken into account like Decision Trees, Random Forests, and XGBoost. When the patient and the doctor are shown the results, the counsellor can talk about any further treatments or medical interventions that might be required.

The process of the Haemogram Report Dataset-Based Automatic Blood Disease Detection System. The report generated by the blood test (haemogram) or pathology test includes the outcomes of entering blood parameters into the input fields of the Automatic Blood Disease Detection System Based on Haemogram Report Dataset. The blood parameter values of XYZ user provided in Table 2.

In compliance with WHO criteria and pathological standards, blood parameter readings are examined, and blood illnesses are determined based on the values of these parameters (high or low).

The blood test (hemogram) results above indicate that anemia was found when the hemoglobin and hematocrit parameter levels were low or decreased from a normal range. The doctor can obtain further guidance for medical therapy, particularly for common or blood disorders, with the use of the Automatic Blood Disease Detection System Based on Haemogram Report Dataset. Furthermore, depending on the circumstances found, the system can suggest a prescription or suggest a course of action.

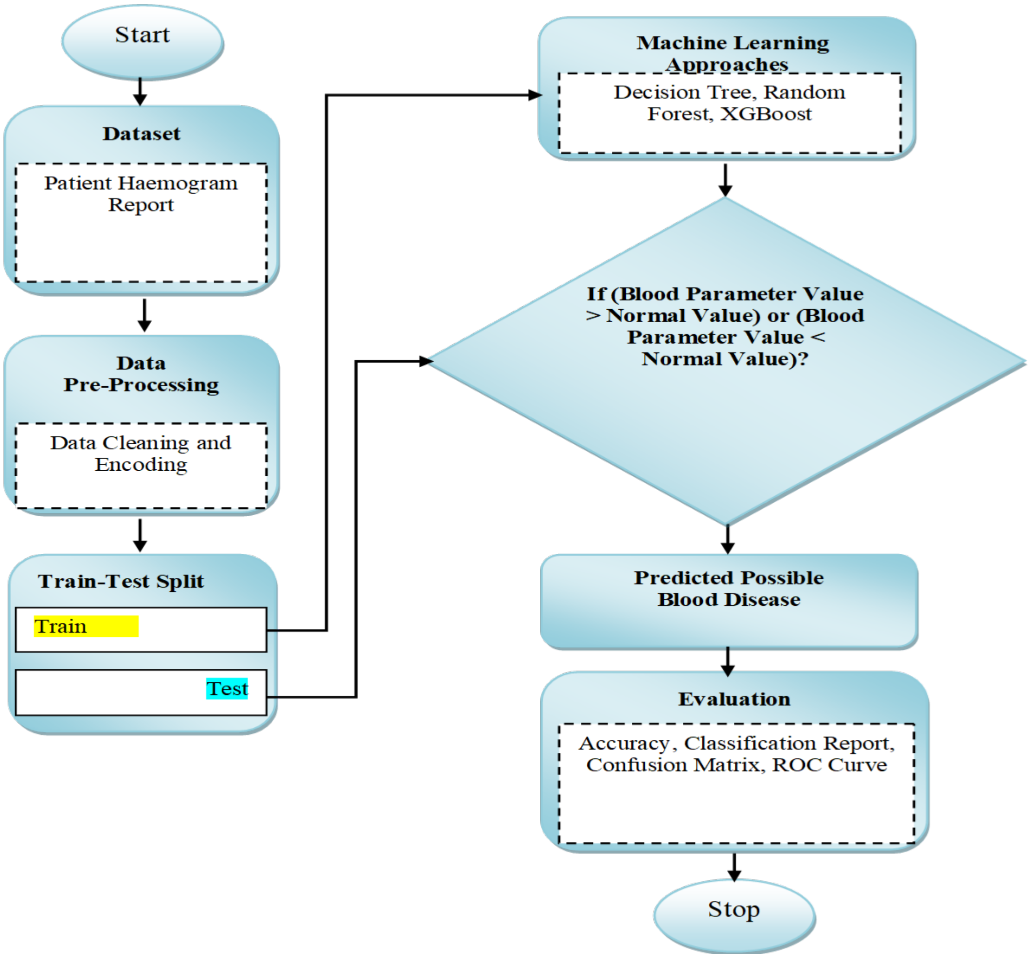

Start
Dataset
Patient Haemogram Report
Data Pre-Processing
Data Cleaning and Encoding
Train-Test Split
Train
Test
Machine Learning Approaches
Decision Tree, Random Forest, XGBoost
If (Blood Parameter Value > Normal Value) or (Blood Parameter Value < Normal Value)?
Predicted Possible Blood Disease
Evaluation
Accuracy, Classification Report, Confusion Matrix, ROC Curve
Stop

Fig 2. Automatic Blood Disease Detection System

Table 2. Blood Test (Haemogram) Report Values Inserted into Automatic Blood Disease Detection System

Blood Parameters	Values	Blood Parameters	Values
WBC	6.6	HCT	29.9
LY	1.8	MCV	75.2
MO	0.5	MCH	22.2
NE	4.3	MCHC	79.6
LY (%)	27.3	RDW	13.3
MO (%)	7.3	PLT	303
NE (%)	65.4	PCT	0.18
RBC	3.96	MPV	9.0
HGB	8.8	PDW	11.8

3.5 Evaluation

Determine the efficacy of the Automatic Blood Disease Detection System Based on Haemogram Report dataset predictive model by using pertinent assessment criteria. Compare their results with those of reliable diagnostic methods in terms of Confusion Matrix, accuracy, recall, precision, f1-score, classification report, and confusion matrix.

Table 3. Confusion Matrix

		Actual Value	
		Positive	**Negative**
Predicted Value	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

$$Accuracy = \frac{True\ Positive\ +\ True\ Negative}{True\ Positive\ + True\ Negative\ +\ False\ Positive\ + False\ Negative} \tag{7}$$

$$Recall = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative} \tag{8}$$

$$Precision = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive} \tag{9}$$

$$F1\ Score\ = 2.\ \frac{Precision\ .\ Recall}{Precision\ +\ Recall} \tag{10}$$

Evaluating accuracy, recall, precision, and f1-score involves using formulas Equations (7), (8), (9) and (10) [24,29]. An assessment categorization report is also utilized to calculate all of the above.

The performance of a binary classification model is evaluated using a graphical representation known as the Receiver Operating Characteristic (ROC) curve [30]. At different threshold values, it compares the True Positive Rate (TPR) versus the False Positive Rate (FPR) in Equations (11) and (12) .

$$TPR = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{11}$$

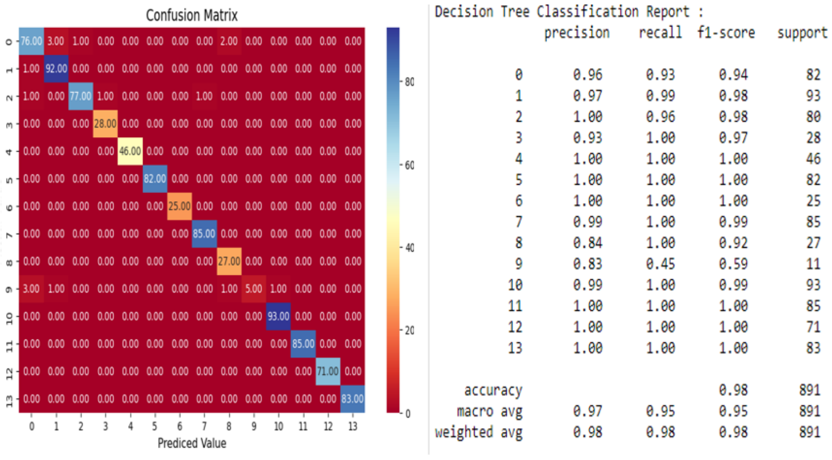
$$FPR = \frac{False\ Positive}{False\ Positive + True\ Negative} \tag{12}$$

An indicator of a binary classification model's effectiveness is the Area Under the Curve (AUC). It alludes to the region beneath the Receiver Operating Characteristic (ROC) curve. Higher values indicate greater model performance. The AUC-ROC value is a numerical number between 0 and 1.

4 Results and Discussion

In this study, we used three machine learning algorithms Decision Tree, Random Forest, and XGBoost to identify particular blood-related issues or diseases using a dataset of 4451 records with 18 distinct blood parameters. We conducted exploratory data analysis to determine the correlations between the attributes, which included correlation matrix analysis and dataset description.

Label encoding was used in this study after the dataset was pre-processed into train and test sets. A random forest classifier was then used to assess the algorithms' performance. Confusion matrices [Figures 3, 4 and 5] and classification reports, which included key performance indicators like accuracy, precision, recall, and F1-score, were used to evaluate the classification models' performance.



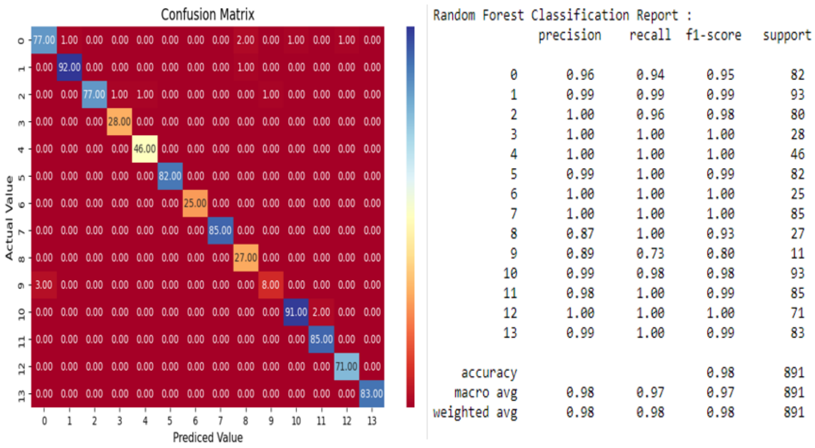
	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	76.00	3.00	1.00	0.00	0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.00	0.00	0.00
1	1.00	92.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	1.00	0.00	77.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	28.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	46.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	82.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.00	25.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	85.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	27.00	0.00	0.00	0.00	0.00	0.00
9	3.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	5.00	1.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	93.00	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	85.00	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	71.00	0.00
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	83.00

```
Decision Tree Classification Report :
              precision    recall  f1-score   support

           0       0.96      0.93      0.94        82
           1       0.97      0.99      0.98        93
           2       1.00      0.96      0.98        80
           3       0.93      1.00      0.97        28
           4       1.00      1.00      1.00        46
           5       1.00      1.00      1.00        82
           6       1.00      1.00      1.00        25
           7       0.99      1.00      0.99        85
           8       0.84      1.00      0.92        27
           9       0.83      0.45      0.59        11
          10       0.99      1.00      0.99        93
          11       1.00      1.00      1.00        85
          12       1.00      1.00      1.00        71
          13       1.00      1.00      1.00        83

    accuracy                           0.98       891
   macro avg       0.97      0.95      0.95       891
weighted avg       0.98      0.98      0.98       891
```

Fig 3. Confusion Matrix and Classification Report of Decision Tree Classifier



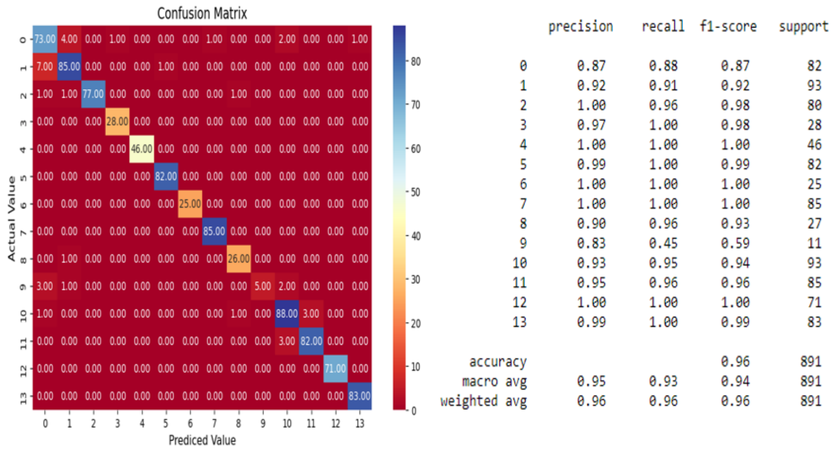
	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	77.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00	0.00	1.00	0.00	1.00	0.00
1	0.00	92.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	77.00	1.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	28.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	46.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	82.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.00	25.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	85.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	27.00	0.00	0.00	0.00	0.00	0.00
9	3.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	91.00	2.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	85.00	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	71.00	0.00
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	83.00

```
Random Forest Classification Report :
              precision    recall  f1-score   support

           0       0.96      0.94      0.95        82
           1       0.99      0.99      0.99        93
           2       1.00      0.96      0.98        80
           3       1.00      1.00      1.00        28
           4       1.00      1.00      1.00        46
           5       0.99      1.00      0.99        82
           6       1.00      1.00      1.00        25
           7       1.00      1.00      1.00        85
           8       0.87      1.00      0.93        27
           9       0.89      0.73      0.80        11
          10       0.99      0.98      0.98        93
          11       0.98      1.00      0.99        85
          12       1.00      1.00      1.00        71
          13       0.99      1.00      0.99        83

    accuracy                           0.98       891
   macro avg       0.98      0.97      0.97       891
weighted avg       0.98      0.98      0.98       891
```

Fig 4. Confusion Matrix and Classification Report of Random Forest Classifier



	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	73.00	4.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	2.00	0.00	0.00	1.00
1	7.00	85.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	1.00	1.00	77.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	28.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	46.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	82.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.00	25.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	85.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	26.00	0.00	0.00	0.00	0.00	0.00
9	3.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.00	2.00	0.00	0.00	0.00
10	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	88.00	3.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	82.00	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	71.00	0.00
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	83.00

```
              precision    recall  f1-score   support

           0       0.87      0.88      0.87        82
           1       0.92      0.91      0.92        93
           2       1.00      0.96      0.98        80
           3       0.97      1.00      0.98        28
           4       1.00      1.00      1.00        46
           5       0.99      1.00      0.99        82
           6       1.00      1.00      1.00        25
           7       1.00      1.00      1.00        85
           8       0.90      0.96      0.93        27
           9       0.83      0.45      0.59        11
          10       0.93      0.95      0.94        93
          11       0.95      0.96      0.96        85
          12       1.00      1.00      1.00        71
          13       0.99      1.00      0.99        83

    accuracy                           0.96       891
   macro avg       0.95      0.93      0.94       891
weighted avg       0.96      0.96      0.96       891
```

Fig 5. Confusion Matrix and Classification Report of XGBoost Classifier

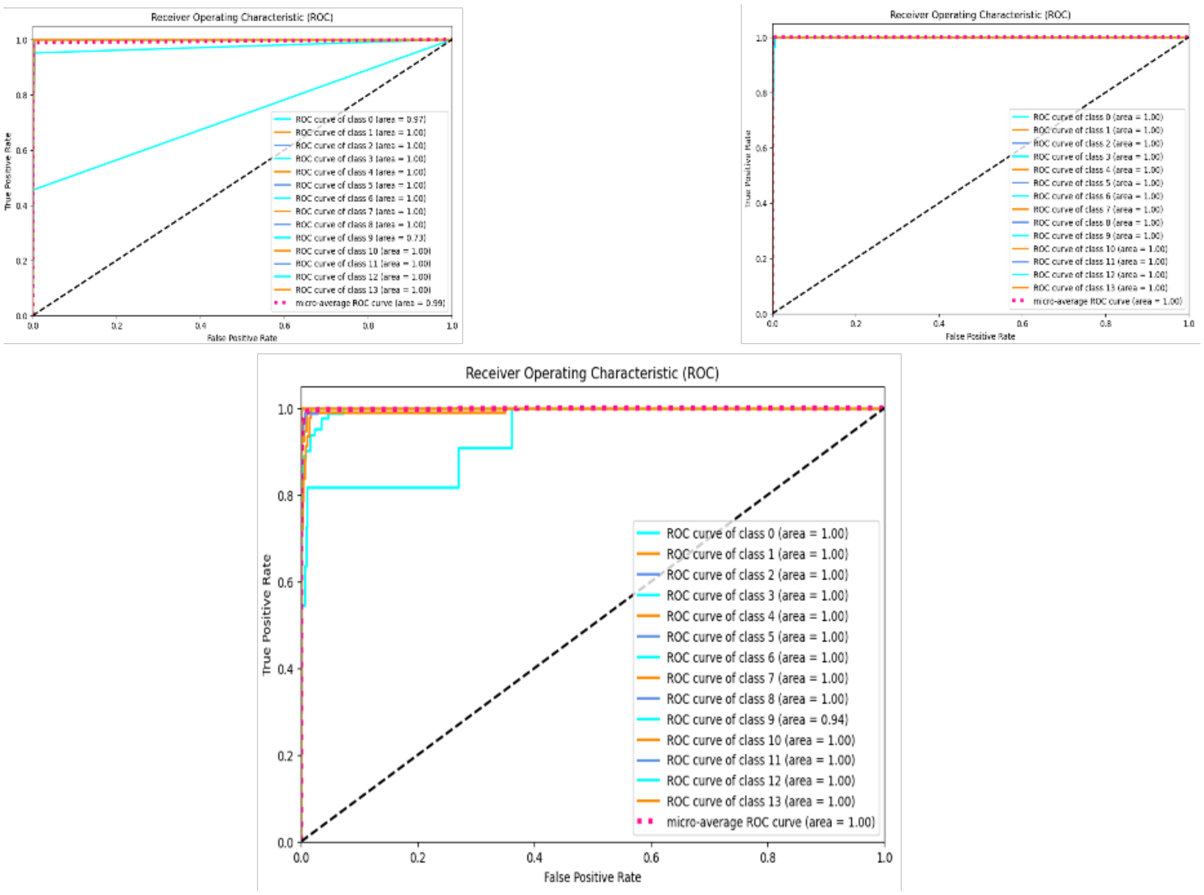


Fig 6. ROC for Decision Tree, Random Forest and XGBoost for Multi-class Dataset

This study used Receiver Operating Characteristic (ROC) curves, an extension of the binary ROC curve, to assess the models' performance in a multi-class classification context. In Figure 6, the ROC curves for each predicted class value using the Decision Tree, Random Forest and XGBoost classifier are displayed.

The findings of this study indicate that the XGBoost model had an accuracy of 96.07%, whilst the Random Forest and Decision Tree classifiers had accuracies of 98.20% and 98.43%, respectively. Table 4 displays each algorithm's accuracy, precision, recall, and F1-score.

Table 4. Illustrated the Accuracy, Precision, Recall and F1-score of Decision Tree, Random Forest and XGBoost

Classifier	Precision	Recall	F1-Score	Accuracy
Random Forest	0.9843	0.9843	0.9843	98.43%
Decision Tree	0.9820	0.9820	0.9820	98.20%
XGBoost	0.9607	0.9607	0.9607	96.07%

Our results show a significant improvement in accuracy compared to earlier research. For instance, research by Parth Verma et al. (2022) found that a Decision Tree classifier could diagnose anemia with 95.45% accuracy[19] and a 95.32% accuracy rate for diagnosing anemia comparable job using a Random Forest classifier[19]. With an accuracy of 96.07% for the XGBoost model, 98.20% for the Decision Tree, and 98.43% for the Random Forest classifier, our study performs better than these earlier reports presented in Table 5.

With an accuracy of 98.43%, our suggested model has shown remarkable performance in identifying a broad range of blood-related conditions, such as Acute or Chronic Blood Loss, Anemia, Autoimmune Disease, Hemolytic Anemia, Immunosuppression, Inflammation, Iron Deficiency, Leukopenia, Lymphoma, Normal, Pernicious Anemia, Sickle Cell Disease, and Stress. The Random Forest method notably obtained an exact accuracy score of 1.0 under these circumstances.

Table 5. Accuracy differentiation between other Model and this Model

Author	Diseases	Dataset	Accuracy
Manish Jaiswal, et. al. (2018) [7]	Anemia	200 with attributes 18	Decision Tree =95.46% Random Forest =95.32%
Mai, C. K.,et. al., (2021) [11]	Anemia	1387 with attributes13	Decision Tree= 95% Random Forest= 96%
Parth Verma, et. al. (2022) [17]	Anemia	200 with attributes 18	Decision Tree =95.45% Random Forest= 95.32%
This Model	Acute or Chronic Blood Loss, Anemia, Iron Deficiency Autoimmune Disease, Hemolytic Anemia, Lymphoma, Pernicious Anemia	4451 with attributes 18	Decision Tree =98.20% Random Forest= 98.43%

Additionally, our investigation has discovered a greater number of blood-related conditions, such as immunosuppression, inflammation, iron deficiency, lymphoma, pernicious anemia, and others, that were not previously studied. This wider spectrum of conditions emphasizes the originality and importance of our study and shows how our methodology may offer a more thorough understanding of blood-related conditions.

Our model has the potential to transform the diagnosis and treatment of blood-related illnesses, as seen by its improved performance and expanded range of illnesses it can identify. Our study's uniqueness and importance are highlighted by a comparison analysis with the body of existing literature, which also underlines the necessity for more research in this field.

Our approach is new in that it applies XGBoost, a potent machine-learning algorithm, to the identification of blood diseases. In comparison to more conventional machine learning algorithms like Random Forest and Decision Tree, our study shows that XGBoost is effective in reaching high accuracy. Accuracy, precision, recall, and F1-score are just a few of the criteria that our study uses to provide a thorough assessment of each algorithm's performance.

In conclusion, research offers a new and efficient method XGBoost for identifying blood-related illnesses, with a high accuracy rate and a wider variety of identifyable disorders. The capacity of our study to offer more thorough insights into blood-related disorders makes it distinctive and a significant contribution to the discipline of hematology. Our findings have important ramifications for the creation of more precise and effective diagnostic instruments for illnesses involving the blood.

5 Conclusion

This study used machine learning algorithms on aberrant blood test findings to generate precise and timely curative analysis for illness therapy. Our study developed hemopathological condition prediction models, with the Random Forest approach obtaining the greatest accuracy of 99.10%. This research offers medical practitioners an effective and trustworthy diagnostic tool. Although our study expands on earlier studies, we are aware of its limitations. Subsequent research endeavors may enhance the precision of alternative machine learning methods, like XGBoost, and investigate the utilization of our models in authentic clinical environments. To improve the results of illness therapy, more studies can look into how our models can be integrated with other diagnostic instruments.

This work offers a unique method for utilizing machine learning algorithms on aberrant blood test data to predict hemopathological diseases. The motivation for this research is the demand for quick and precise illness diagnosis. A vital medical test called a total blood count can identify several signs of illness. Our goal is to overcome the shortcomings of earlier research by developing models for hemopathological condition prediction.

With an accuracy of 99.10%, the Random Forest approach is superior to other algorithms; Decision Tree classifiers come in second with 98.87% accuracy, while XGBoost comes in third with 96.63% accuracy. Our research indicates increased prediction accuracy for blood-related disorders when compared to earlier studies. As a trustworthy and effective diagnostic tool for medical practitioners, our results meet the goals of this study. The most useful algorithm for forecasting hemopathological situations is the Random Forest approach, which provides readers with a fresh piece of knowledge. Our study offers specific advice on how medical practitioners should use this method for diagnosing illnesses.

References

1) Alsheref FK, Gomaa WH. Blood Diseases Detection using Classical Machine Learning Algorithms. *International Journal of Advanced Computer Science and Applications*. 2019;10(7):1–5. Available from: https://thesai.org/Downloads/Volume10No7/Paper_12-Blood_Diseases_Detection_using_Classical_Machine.pdf.
2) DGaikwad, Mahale V, Gaikwad AT. A Review on Blood Disease Detection using Artificial Intelligence Techniques. In: 2024 IEEE International Conference on Big Data & Machine Learning (ICBDML). IEEE. 2024. Available from: https://www.semanticscholar.org/paper/A-Review-on-Blood-Disease-Detection-using-Gaikwad-Mahale/c103d77886242e1ccad15ab61ac604e80b7b8035.
3) Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *Jama* . 2017;6:517–518. Available from: https://jamanetwork.com/journals/jama/article-abstract/2645762.
4) Nithya B, Ilango V. Predictive analytics in health care using machine learning tools and techniques. In: and others, editor. International Conference on Intelligent Computing and Control Systems (ICICCS), 15-16 June 2017, Madurai, India. IEEE. 2018. Available from: https://ieeexplore.ieee.org/abstract/document/8250771.
5) Gunčar G, Kukar M, Notar M, Brvar M, Černelč P, Notar M, et al. An application of machine learning to haematological diagnosis. *Scientific reports* . 2018;8:1–12. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC5765139/pdf/41598_2017_Article_18564.pdf.
6) Soni M, Varma DS. Diabetes Prediction using Machine Learning Techniques. *International Journal of Engineering Research & Technology (IJERT)*. 2020;9(9). Available from: https://www.ijert.org/diabetes-prediction-using-machine-learning-techniques.
7) Abdurrahman G, Sintawati M. Implementation of xgboost for classification of parkinson's disease. In: 3rd International Conference on Combinatorics, Graph Theory, and Network Topology, 26-27 October 2019, East Java, Indonesia;vol. 1538. IOP Publishing Ltd. 2020. Available from: https://iopscience.iop.org/article/10.1088/1742-6596/1538/1/012024/pdf.
8) Fahim NW, Eshti SA, Nura KA, Abir MJH, Mahbub MNI. Parkinson Disease Detection: Using XGBoost Algorithm to Detect Early Onset Parkinson Disease. 2020. Available from: https://www.researchgate.net/publication/354835987_Parkinson_Disease_Detection_Using_XGBoost_Algorithm_to_Detect_Early_Onset_Parkinson_Disease.
9) Farajollahi B, Mehmannavaz M, Mehrjoo H, Moghbeli F, Sayadi MJ. Diabetes diagnosis using machine learning. *Frontiers in Health Informatics*. 2021;10(1):1–5. Available from: https://www.researchgate.net/publication/350745659_Diabetes_Diagnosis_Using_Machine_Learning.
10) Mai CK, Reddy AB, Raju KS. Algorithms for Intelligent Systems. In: and others, editor. Machine Learning Technologies and Applications. Springer. 2021. Available from: https://link.springer.com/book/10.1007/978-981-33-4046-6.
11) Llaha O, Rista A. Prediction and Detection of Diabetes using Machine Learning. In: and others, editor. Proceedings of the 4th International Conference on Recent Trends and Applications in Computer Science and Information Technology;vol. 2021. ;p. 1–9. Available from: https://ceur-ws.org/Vol-2872/paper13.pdf.
12) Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. *ICT Express*. 2021;7(4):432–439. Available from: https://doi.org/10.1016/j.icte.2021.02.004.
13) Jackins V, Vimal S, Kaliappan M, Lee MY. AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *The Journal of Supercomputing*;77:5198–5219. Available from: https://link.springer.com/article/10.1007/s11227-020-03481-x.
14) Akter L, Ferdib-Al-Islam. Dementia Identification for Diagnosing Alzheimer's Disease using XGBoost Algorithm. In: Proceedings of the 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD);vol. 2021. IEEE. ;p. 205–209. Available from: https://ieeexplore.ieee.org/document/9396777.
15) Jadhav N, Phad S, Kamble S. Detecting Parkinson's Disease Using Xgboost Classifier Machine Learning. *International Research Journal of Modernization in Engineering Technology and Science*. 2021;3(5):1–5. Available from: https://www.irjmets.com/uploadedfiles/paper/volume3/issue_5_may_2021/9668/1628083388.pdf.
16) Verma P, Chopra V. Machine Learning Algorithms for Anemia Disease Prediction - A Review. *International Research Journal of Modernization in Engineering Technology and Science*. 2022;4(4):1–5. Available from: https://www.irjmets.com/uploadedfiles/paper/issue_4_april_2022/21669/final/fin_irjmets1651552586.pdf.
17) Boukhatem C, Youssef HY, Nassif AB. Heart disease prediction using machine learning. *Proceedings of the IEEE Advances in Science and Engineering Technology International Conferences (ASET-2022)*;2022:1–6. Available from: https://ieeexplore.ieee.org/document/9734880.
18) Santos DH. Parkinson's Disease Detection using XGBoost and Machine Learning. *medRxiv preprint*. 2023. Available from: https://www.medrxiv.org/content/10.1101/2023.10.23.23297369v1.full-text.
19) Budholiya K, Shrivastava SK, Sharma V. An optimized XGBoost based diagnostic system for effective prediction of heart disease. *Journal of King Saud University - Computer and Information Sciences*. 2022;34(7):4514–4523. Available from: https://doi.org/10.1016/j.jksuci.2020.10.013.
20) Doki S, Devella S, Tallam S, Gangannagari SSR, Reddy PS, Reddy GP. Heart Disease Prediction Using XGBoost. *Proceedings of the IEEE third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT) 2022*. 2022. Available from: https://ieeexplore.ieee.org/document/9917678.
21) Gaikwad D, Mahale V, Gaikwad A, et al. A Review on Blood Disease Detection using Artificial Intelligence Techniques. *2024 IEEE International Conference on Big Data & Machine Learning (ICBDML)*. 2024;p. 21–26. Available from: https://doi.org/10.1109/ICBDML60909.2024.10577320.
22) Gaikwad DK, Gaikwad A. Disease Prediction using Artificial Intelligence Techniques: A Review. *Industrial Engineering Journal*. 2023;XVI(3):92–103.
23) Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*. 2021;2. Available from: https://link.springer.com/article/10.1007/s42979-021-00592-x.
24) Sah S. Machine Learning: A Review of Learning Types. *Preprints*. 2020. Available from: https://www.researchgate.net/publication/342890321_Machine_Learning_A_Review_of_Learning_Types.
25) Machine Learning Tutorial. . Available from: https://www.javatpoint.com/machine-learning.
26) Gaikwad DK, Gaikwad A, AGaikwad. Blood Disease Detection System Based on Haemogram Report using Decision Tree Algorithm. In: and others, editor. Proceedings of the NIELIT's International Conference on Communication, Electronics and Digital Technology;vol. 1023. 2024. Available from: https://link.springer.com/chapter/10.1007/978-981-97-3604-1_15.
27) Salim NOM, Abdulazeez AM. Human Diseases Detection Based On Machine Learning Algorithms: A Review. *International Journal of Science and Business*. 2021;5(2):102–113. Available from: https://ideas.repec.org/a/aif/journl/v5y2021i2p102-113.html.
28) Abdul-Jabbar SS, Farhan AK. Complete Blood Count (CBC) Test. . Available from: https://www.kaggle.com/datasets/ahmedelsayedtaha/complete-blood-count-cbc-test.

29) Gaikwad DK, Gaikwad A. Pandemic Predictor Pro using Machine Learning Algorithms. *Indian Journal of Natural Sciences*. 2024;14(82):68306–68314. Available from: https://scholar.google.co.in/citations?view_op=view_citation&hl=en&user=gkSEVgwAAAAJ&citation_for_view=gkSEVgwAAAAJ:Y0pCki6q_DkC.

30) Centers for Disease Control and Prevention. Government agency. 2007. Available from: http://www.cdc.gov/nchs/data/nhanes/nhanes_05_06/cbc_d_met.pdf.