Analyzing and Predicting Anaemia with Advanced Machine Learning Techniques with Comparative Analysis

Sandeep Kumar Chundru
University of Central Missouri

Mukund Sai Vikram Tyagadurgam
University of Illinois at Springfield

Venkataswamy Naidu Gangineni
University of Madras, Chennai

Sriram Pabbineedi
University of Central Missouri

Ajay Babu Kakani
Wright State University

Sri Krishna Kireeti Nandiraju
University of Illinois at Springfield

ABSTRACT

The rising prevalence of anaemia poses significant health challenges globally, necessitating accurate and timely diagnostic methods. This work applies the AdaBoost classification model to a Complete Blood Count (CBC) dataset to provide a reliable machine learning method for anaemia prediction. The methodology used consists of extensive pre-processing including data cleaning, one-hot encoding, z score normalization and automated feature selection to preserve the data's integrity and make the model simpler. The experimental results proved that the presented AdaBoost model achieved a promising accuracy of 92.7%, good precision of 84%, fair recall of 94% and good F1 value of 88.7%, indicating effectively well-balanced classification performance. Further, ROC curve analysis, with AUC of 0.94, confirms superior discriminatory capability. It compared their results to existing models, LogitBoost and Random Forest (RF); LogitBoost yields an accuracy slightly low at 89.3% and RF at 67.1%. The results highlight the capabilities of the AdaBoost model for early, accurate anaemia detection, providing substantial improvements over conventional diagnostic measures and improving clinical decision making in real time healthcare applications.

Keywords

Anaemia Detection, AdaBoost, CBC Dataset, Diagnostic Accuracy, Predictive Analytics, Health Informatics.

1. INTRODUCTION

Data-driven technologies are being used increasingly by healthcare systems around the world to improve patient diagnosis, treatment and outcomes. Approximately 1.6 billion individuals worldwide suffer from anaemia, making it one of the most prevalent public health conditions [1]. The inability for the body to carry oxygen properly due to red blood cell or hemoglobin deficiency leads to tiredness, weakness and when this is severe, life threatening problems. Women, children, and the elderly are more likely to get anaemia than men and people with lower or middle incomes [2][3]. Anaemia is a medical disease when blood does not contain adequate red blood cells or hemoglobin. It inhibits the oxygen getting down to the cells. Common signs of deeper health problems include feeling tired, feeling weak, pale skin and out of breath.

Traditional diagnostic methods for anaemia typically rely on laboratory tests such as complete blood counts (CBC) and hemoglobin level assessments. While effective, these tests can be time-consuming, costly, and inaccessible in resource-limited settings [4][5]. Moreover, they often do not leverage the full potential of the available clinical and demographic data to predict anaemia risk. As a result, many cases go undetected until symptoms become severe. More effective and scalable methods are needed right away to improve early detection and help healthcare professionals make quick choices [6]. The most common risk factor for anaemia around the world is not getting enough iron because of poor diet, problems absorbing iron, or both. This makes the body make fewer red blood cells. Anaemia can also be caused by genetic conditions like thalassemia, short-term and long-term infections (which cause inflammation), and not getting enough micronutrients (like vitamin A, vitamin B12, and folate). A clinical sign of anaemia is the amount of hemoglobin in the blood.

ML, a branch of AI, has shown great promise in the healthcare sector for predictive analytics and decision support [7]. ML models can look through huge amounts of organized and unstructured data to find patterns, correlations, and risk factors that aren't obvious. In recent years, ML has been successfully applied in areas such as cancer detection, cardiovascular risk prediction, and diabetes management [8][9]. However, the application of ML to anaemia detection and analysis remains relatively underexplored, particularly from a comparative perspective involving multiple algorithms. A comprehensive ml-based framework for the prediction of anaemia, with a focus on comparative performance analysis. By combining clinical data with advanced analytics, their work aims to support the development of intelligent diagnostic tools that can be integrated into routine healthcare practices [10].

1.1 Motivation and Contribution of Study

The goal of this research is that anaemia is becoming more common and reliable and quick diagnostic tools are very important in healthcare. Traditional ways of diagnosing often depend on biased evaluations, which can lead to wrong diagnoses or delayed treatment, both of which are bad for patients. One possible strategy to increase diagnosis accuracy and facilitate early management is to include ML algorithms into the anaemia detection process. A Complete Blood Count (CBC) dataset is used in this study to improve the classification of anaemia using advanced ml techniques. The AdaBoost model is especially used in this study. These are the most important things that this study adds:

- The study uses the AdaBoost classification model to make disease prediction more accurate based on CBC data. This leads to more accurate and early diagnoses in hospital settings.
- A comprehensive data pre-processing framework including missing value handling, outlier treatment ensures data quality and integrity, which is critical for achieving high model performance.
- Using feature selection techniques helps find the most important clinical indicators, which lowers the complexity of the model while keeping or better classification performance.
- The research uses several rating methods (accuracy, precision, F1-score, as well as recall) to give a full picture of the model's prediction skills. This offers more details about how useful it is in real life.

1.2 Novelty and Justification of the Paper

This paper's originality is found in its application of a robust ensemble learning technique AdaBoost for the accurate classification of anaemia using Complete Blood Count (CBC) data, a widely available yet underutilized diagnostic resource. Unlike traditional diagnostic methods that often rely on singular clinical indicators, this study integrates multiple hematological parameters with ml to enhance predictive accuracy and reliability. The justification for this approach stems from the growing need for cost-effective, data-driven tools in resource-constrained healthcare environments. Utilizing AdaBoost's ability to handle large, uneven datasets and enhance classification performance, the suggested method provides a scalable way to find early signs of anaemia, enabling timely medical intervention and better patient outcomes.

1.3 Structure of the Paper

The paper is formatted as follows: In Section 2, related research on ml methods for classifying anaemia is looked at. In Section 3, the method that initiated this thesis is described, including data collection, preparation and implementation of the model. Within Section 4, it sees the AdaBoost model's data and performance report. Section 5 ends with some ideas and thoughts about where more research should be done in the future.

2. LITERATURE REVIEW

In this section, the paper explores progress in detecting and classifying anaemia with a focus on the role of ml techniques in enhancing diagnostic accuracy for Complete Blood Count (CBC) data. This thesis emphasizes the potential of these advanced methodologies to improve the detection of anemia, a common health problem. Here are some review works of some key:

Hortinela et al. (2019) RBC is looked at using standard microscope which can result to wrong results and a lot of work for the experts. A Raspberry Pi-based device that can look at the RBC's area, shape geometric factor (SGF), perimeter, diameter, and the suggestion is to locate the target flag and central pallor. This will help medical technicians, hematologists, and pathologists identify RBC. Multiple approaches were used in past studies on the same subject. Utilizing an Artificial Neural Network (ANN), one study correctly identified RBC with 90.54% accuracy. The accuracy of another study, which used a radial basis function network, was 83.3% [11].

Hafeel et al. (2019) intend to create a tool that can identify anaemia, a disease that happens when the body doesn't have enough Fe3+ ions. With this disease, organs could stop working, which could lead to a heart attack or death. When someone has anaemia, their blood's redness decreases. Using this information, the server will decide if the person has anaemia or not. It has an accuracy rate of almost 83% when tested on the anaemia patient, and it meets the criteria for an accurate result when compared to the real tests that were done [12].

Gebreweld and Tsegaye (2018) Sociodemographic and clinical information about the people who took part in the study was gathered through interviews and reviews of their medical records using a structured questionnaire that had already been tried. A P number of less than 0.05 was always thought to be statistically significant. Outcome. 11.6% of people (95% CI: 7.8%–14.8%) were found to have anaemia. Anaemia was more common in pregnant women in the second and third trimesters than in those in the first trimester (AOR (95% CI), 6.72 (1.17–38.45; P=0.03). Compared to pregnant women who took iron/folic acid supplements, those who did not had a higher risk of anaemia, as indicated by the AOR (95%CI) [13].

Roychowdhury et al. (2017) to check for anaemia by looking at the pale areas of the conjunctiva and mouth. The sclera and the conjunctivae are automatically split into areas of interest on the eye pallor site photos. Then, color-plane-based feature extraction is used to narrow down the set of features. ml methods are then employed to grade the images for anaemia. Picture-level classification algorithms are used in this study to show whether an image is normal (class 0), pale (class 1), or has other problems (class 2). Eye pallor site images can be read 86% correctly, 85% precisely, and 67% reliably with the suggested method [14].

Hennek et al. (2016) describes how aqueous multiphase systems (AMPS) can be used to quickly and affordably determine whether a person has IDA. These are salt and biodegradable polymer mixes that are thermodynamically stable and generate distinct layers with abrupt density shifts. Only two minutes were spent centrifuging the tests (n = 152) in an inexpensive centrifuge. After being visually read, the area under the curve (AUC) that they displayed was 0.89, indicating that they were 78% specific (68–86%) and 84% sensitive (72–93%). The AMPS test is comparable to clinical tests such as reticulocyte hemoglobin concentration (AUC = 0.9) and performs better than merely examining hemoglobin (AUC = 0.73). Regular ML techniques were used to examine the test images that were taken by a regular desktop scanner. The IDA diagnosis was somewhat improved (sensitivity of 90%, specificity of 77%, and range of 83 to 96%), and it also helped forecast other crucial characteristics of red blood cells, like the average composition of corpuscular hemoglobin [15].

Villagrasa et al. (2015) Electronics that are made to use little power rely on simple standards. This makes for a durable and low-power device that can be taken anywhere and has a long battery life. 50 µL blood samples are detected using a three-gold electrode sensor. It is disposable, cheap, and doesn't need to be marked. The device was tested on 24 blood samples from four hospitalized patients with anaemia and was able to identify the condition. This means that the small point-of-care device responded, worked well, and could be counted on to find anaemia. 2.83 percent of that number was different from the mean, which was 2.57 percent. Not a single case was above 5% [16].

The comparative analysis of background study based on their methodology, data, key findings, limitations, and future work is given in Table 1.

Table 1. Summary of Reviewed Works on Anaemia Prediction using Machine Learning techniques.

Author	Methodology	Data	Key Findings	Limitations and Future Work
Hortinela et al. (2019)	Raspberry Pi device using image processing to analyze RBC shape and structure	RBC microscopy images	Helps medical staff by automatically identifying RBC parameters like area, SGF, central pallor, etc. Prior ANN model achieved 90.54% accuracy; RBFN achieved 83.3%.	Lack of details on clinical testing or real-time deployment. Future work can include larger datasets and real-world validation.
Hafeel et al. (2019)	Portable device measuring blood's red intensity + questionnaire, analyzed via server	Blood color intensity + questionnaire	Achieved 83% accuracy in anaemia prediction using combined sensor + app inputs.	Red intensity measurements may be inconsistent due to lighting/skin tone. Future work can improve robustness and validation across demographics.
Gebreweld and Tsegaye (2018)	Logistic regression on clinical and demographic data	CBC + questionnaire from pregnant women	Found 11.6% anaemia prevalence; higher risk in 2nd/3rd trimesters and in those without supplements.	Study limited to pregnant women in one region. Broader sampling and inclusion of nutritional data could enhance insights.
Roychowdhury et al. (2017)	ML-based classification using image segmentation of conjunctiva and tongue pallor	Eye and tongue	Eye pallor analysis achieved 86% accuracy, 85% precision. Demonstrated viability of image-based screening.	Recall was only 67%. Performance may vary with lighting and image quality. Future work can focus on improving recall and automating preprocessing.
Hennek et al. (2016)	AMPS in capillary tubes, analyzed visually and via ML on scanned test images	152 fingerstick blood samples	Achieved 90% sensitivity, 77% specificity for IDA diagnosis. AUC of 0.89. Outperformed Hb-only tests.	Manual reading required; though ML helps, more automation needed. Broader clinical testing would strengthen validity.
Punter-Villagrasa et al. (2015)	Portable low-power electrochemical sensing device	24 blood samples from 4 patients	Accurate (2.83% error), mobile, and energy-efficient anaemia detection. Robust performance.	Generalization is limited by a small sample size. Future studies could optimize for more situations and increase the sample size.

3. METHODOLOGY

The method starts with getting the CBC (Complete Blood Count) information. Next, extra columns are taken out, missing values are dealt with, and outliers are dealt with. Preprocessed data is encoded using One-Hot Encoding, and target labels are made. Relevant features are chosen, and the data is separated into training and testing groups after preprocessing. The AdaBoost classification model is then taught with the training group. The model's accuracy, F1-score, recall and precision are used to evaluate its performance. Lastly, the success metrics are used to put together and analyze the results. For correct classification using the CBC data, this method guarantees strong model training and reliable evaluation. Using AdaBoost improves prediction even more by mixing several weak learners into a strong classifier. Figure 1 illustrates the proposed workflow for anaemia detection using CBC data, aimed at enhancing diagnostic capabilities in clinical settings.

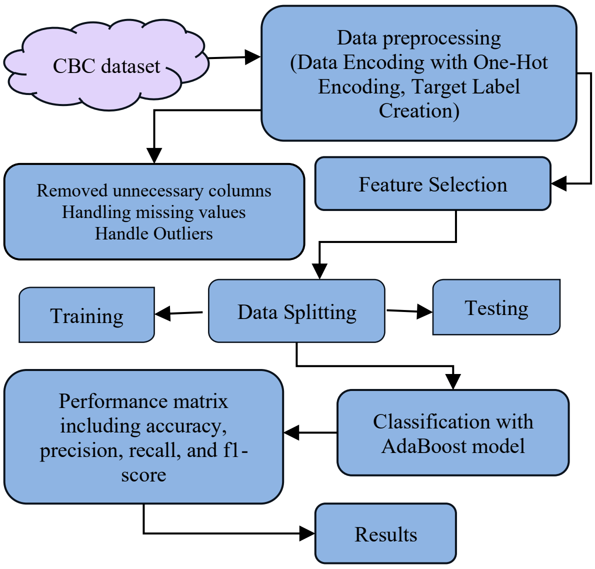
CBC dataset
Data preprocessing (Data Encoding with One-Hot Encoding, Target Label Creation)
Removed unnecessary columns
Handling missing values
Handle Outliers
Feature Selection
Training
Data Splitting
Testing
Performance matrix including accuracy, precision, recall, and f1-score
Classification with AdaBoost model
Results

Fig 1: Data Flowchart Diagram for Anaemia detection

A brief description of each phase in a data flow diagram is provided below:

3.1 Data Collection

The Complete Blood Count (CBC) sample that was used in this study came from Kaggle and has the results of CBC tests that were done at the Eureka Diagnostic Centre in Lucknow, India. After excluding pregnant women and individuals under 15, 364 adult male and female samples remained. All tests were performed using standard hematology analyzer protocols to ensure data accuracy and consistency. Some visualization is here.

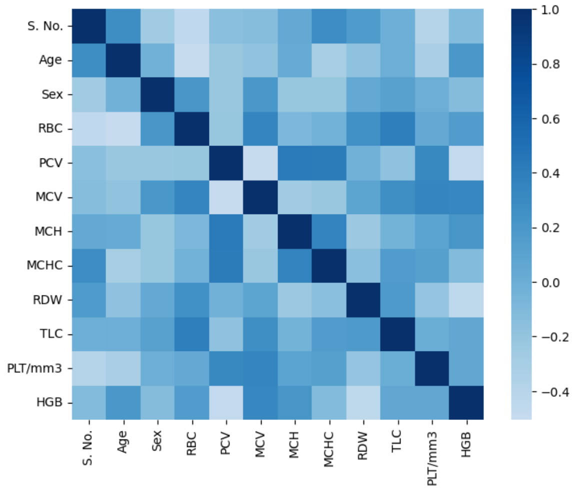

Fig 2: The Correlation Coefficient Among the Variables

Figure 2 presents a correlation heatmap showing the relationship strength between hematological and demographic variables. Darker shades represent stronger correlations. Positive and negative associations are visualized, helping identify multicollinearity or independent features.

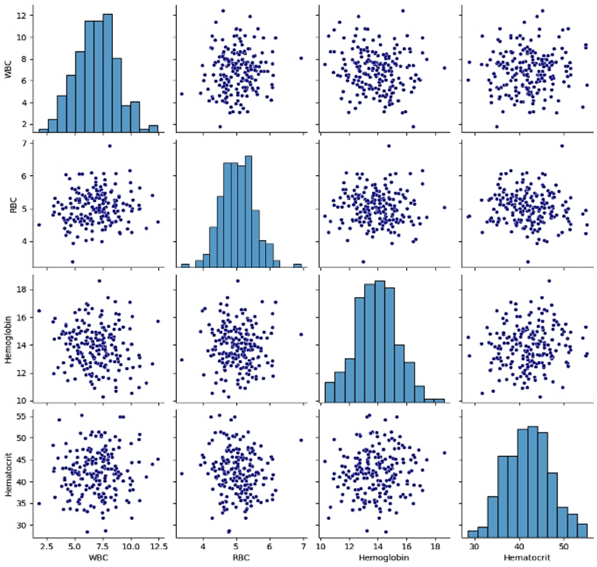

Fig 3: Pair Plot of Hematological Parameters

Figure 3 shows a pair plot visualizing relationship between WBC, RBC, Hemoglobin, and Hematocrit. Individual histograms, which show distributions, are shown on diagonal maps. Off-diagonal scatter plots show patterns of association. For example, RBC and Hemoglobin seem to be linked in a good way. The symmetric layout aids in spotting trends and potential linear associations across features.

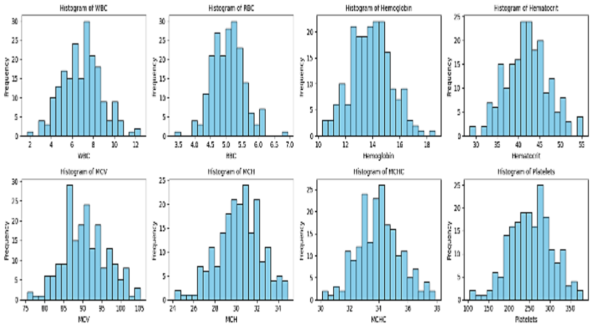

Fig 4: Distribution of Hematological Features

Figure 4 displays histograms for eight hematological features: WBC, RBC, MCV, MCH, MCHC, hemoglobin, hematocrit, and platelets. Each plot shows the distribution pattern of values, revealing symmetry, spread, and potential outliers. Most variables exhibit approximately normal distributions, essential for understanding data variability and guiding further statistical or ml analysis.

3.2 Data Preprocessing

The first and most important step in turning raw data into useful information is data preparation. In general, raw data might not be full, be repeated, or be noisy. In this study initially, the raw Complete Blood Count (CBC) data is cleaned to eliminate duplicates and irrelevant entries. This careful preparation makes the data better and gets it ready for feature selection and model training, which makes the anaemia detection system work better overall. Here are the main steps in editing:

- **Removed Unnecessary Columns:** To make the dataset more streamlined, unnecessary columns were eliminated, enhancing its relevance and improving the analysis process.
- **Handling Missing Values:** Missing data patterns show which numbers are missing from a set of data. But there isn't a standard list of data trends that are missing.
- **Handle Outliers:** The goal of outlier identification is to find a set of data points that are not like the other data points in a dataset. Techniques for finding outliers can be done numerically or graphically, based on the need. The Z-score is the next simple statistical tool that can be used to find outliers based on a certain trait. This number tells you how far the data point or sample's value is from the feature's mean. In this way, the Z-score is shown in huation (1):

$$z - score = \frac{x_i - \tilde{x}}{\sigma} \quad (1)$$

the value of feature x for the its sample, x_i, and the standard deviation σ and mean $\tilde{x}$ of the range of feature x are given by

3.3 Data Encoding with One-Hot Encoding

Data encoding techniques are intended to reduce power consumption caused by changes in the chip's connections. As shown in the One-Hot encoding, a character or word is represented by a vector with one element that is one and all the other elements being one. When mapping a text to x, its length is equal to the vocabulary's size, and the part that is set to one depends on where it is in the vocabulary. The formula is shown in Equations (2) or (3).

$$X_i = [x_1, x_2, \dots, x_n] \tag{2}$$

Where,

$$x_j = \begin{cases} 1, \; if\; j = i \\ 0, \; otherwise \end{cases} \tag{3}$$

The encoded binary value x_j represents category j, where i is the index of the categorical value among n unique categories. The encoding assigns a binary vector of length n, setting the position corresponding to the category to 1, while all other positions remain 0.

3.4 Target Label Creation (Anaemia Classification)

Target labels for anaemia classification were created based on hemoglobin levels, categorizing individuals into anemic and non-anemic groups. This clear labelling makes sure that the model is trained correctly and makes ml algorithms better at predicting anaemia.

3.5 Feature Selection

In this research, feature selection was used to find the most important factors that affect the diagnosis of anaemia. "Age," "Sex," and "Hb" (Hemoglobin) amounts were chosen as the most important factors because they have a direct effect on diagnosing anaemia. Features that weren't needed or were used more than once were taken away to make the model work better and be simpler. The choice of factors makes sure that the predictive model focusses on the most important ones for correctly classifying anaemia.

3.6 Data Splitting

Splitting data into groups that can be used for training and testing is called data splitting. 80% of the data is used to train the machine learning model, with the remaining 20% retained to assess the model's performance on untested data.

3.7 Proposed Model of AdaBoost Model

The AdaBoost classification method is very good at what it does and doesn't make mistakes when it applies to other situations. use SVM, CART, and C4.5 for AdaBoost if you are a weak student. The goal of AdaBoost is to make a group of weak algorithms stronger. By repeating steps more than once, AdaBoost lowers the number of classification errors [17]. With this method, the weight of the weak classifier with a high classification error rate is increased, while the weight of the weak classifier with a low classification error rate is decreased, until the round number or error rate is found. If you're learning style is weak, choose the choice tree [18]. T is the number of weak learners shown by ht(x) Equation (4) is used to figure it out.

$$\theta_t = \frac{1}{2} In \frac{1-\epsilon_t}{\epsilon_t} \tag{4}$$

ϵ_t is the rate of classification errors in every weak learner. Using Equation (5), Each data point's weight is modified.

$$\omega_{t+1}(x_i, y_i) = \frac{\omega_t(x_i,y_i)+e^{-\theta_t y_i f_{t(x_i)}}}{z_t} \tag{5}$$

where $\omega_t$is the data point's weight and z_t is a normalization factor that makes sure the sum of the weights of all the instances is 1. There is a strong classifier and a weak student. As shown in Equation (6), the linear combination of the basic classifications is:

$$H(x) = \sum_{t=1}^{T} \theta_t h_t(x) \tag{6}$$

The final prediction is the weighted sum of all the classifiers forecasts, which can be shown in concert as Equation (7):

$$G(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x)) \tag{7}$$

The AdaBoost model final hypothesis G(x) which consists of weighted summation of weak classifiers $h_t(x)$ with their weights, α_t, followed by a sign function

3.8 Performance Matrix

A Performance Matrix is a set of metrics that one figures out to see how good a ml model works. Confusion matrix components include False Positive (FP), True Negative (TN), False Negative (FN), and True Positive (TP). This is an example of a confusion matrix. It shows how forecasts turned out. This tells you how many correct and incorrect answers a class has had. In order to assess the effectiveness of different categorization techniques, metrics such as F1 score, precision, accuracy, recall and others are used, along with a confusion matrix. The most important factors are:

- **TP:** Correctly predicted positive observations
- **TN:** Correctly predicted negative observations
- **FP:** Incorrectly predicted positive observations
- **FN:** Incorrectly predicted negative observations.

3.8.1 Accuracy

The dataset says that accuracy is the number of correctly forecast observations. To find out, divide the total number of predictions by the number of correct possibilities. The given method is shown in Equation (8).

$$\text{Accuracy} = \frac{\text{TN + TP}}{\text{TP + TN + FP + FN}} \tag{8}$$

3.8.2 Precision

The precision of a classifier tells how well it predicts good samples. Divide the total number of samples that were expected to be positive by the number of actual positives, and you get the answer. It is found mathematically using Equation (9):

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{9}$$

3.8.3 Recall

Recall or sensitivity is the percentage of correctly anticipated positive cases in the dataset compared to all actual positive cases. One way to compute it is to divide the total number of inaccurate negative forecasts by the number of accurate positive forecasts. Equation (10) provides a mathematical representation of it:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{10}$$

3.8.4 F1-Score

The F-measure, which is the harmonic mean of accuracy and recall, finds a balance between the two. See below for the following Equation (11):

$$\mathrm{F1} = \frac{2*(\mathrm{precision}*\mathrm{recall})}{\mathrm{precision}+\mathrm{recall}} \quad (11)$$

3.8.5 ROC

The Receiving Operating Characteristics (ROC) graph displays how well a classification model works at various decision levels, such as the False Positive Rate (FPR) and True Positive Rate. What follows are Equations (12 and 13).

$$FPR = \frac{FP}{TN+FP} \quad (12)$$

$$TPR = \frac{TP}{TP+FN} \quad (13)$$

These performance matrices are utilized for comparative analysis and to evaluate the model performance for the Complete Blood Count (CBC) dataset.

4. RESULTS ANALYSIS AND DISCUSSION

This study assesses the performance of the AdaBoost model applied to Anaemia Prediction. Experiments were conducted using Python 3 with TensorFlow and Scikit-Learn on a 64-bit Windows 10 system featuring an Intel i7 processor (3.60 GHz, four-core) and 16 GB RAM. The AdaBoost model's performance metrics are provided in Table 2, indicate a commendable accuracy of 92.7%, ensuring reliable overall classification. The precision of 84% suggests a moderate rate of false positives, while the recall of 94% demonstrates its strong capability in detecting actual Anaemia. The model is effective, as seen by its balanced accuracy and recall performance, as shown by its F1-score of 88.7%. The following results of the proposed model are discussed below:

Table 2. Performance Metrics of the Proposed AdaBoost Model for Anaemia Prediction

Performance Metric	AdaBoost Model
Accuracy	92.7
Precision	84
Recall	94
F1-score	88.7

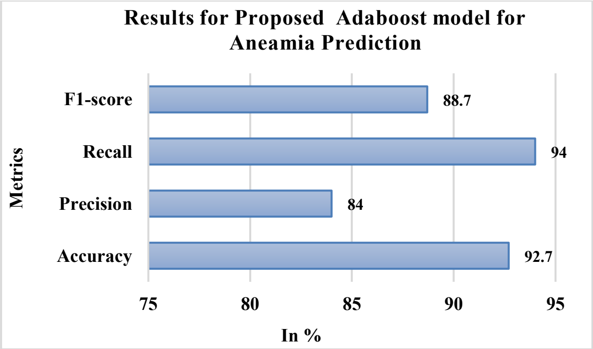

Fig 5: Results for proposed Adaboost model for Aneamia prediction

Table 2 and Figure 5 presents the performance metrics of the proposed AdaBoost model for anaemia prediction, demonstrating its strong predictive capability. The model achieved an accuracy of 92.7%, indicating a high overall correctness in classification. With a precision of 84%, the model shows a good ability to correctly identify true anaemia cases while minimizing false positives. Its recall of 94% highlights the model's effectiveness in detecting the majority of actual anaemia cases, reducing the likelihood of missed diagnoses. The F1-score of 88.7% provides a balanced measure of precision and recall, confirming that the model maintains both sensitivity and reliability in predictions. These results collectively suggest that AdaBoost is a robust approach for anaemia detection, offering a reliable diagnostic aid in clinical applications.

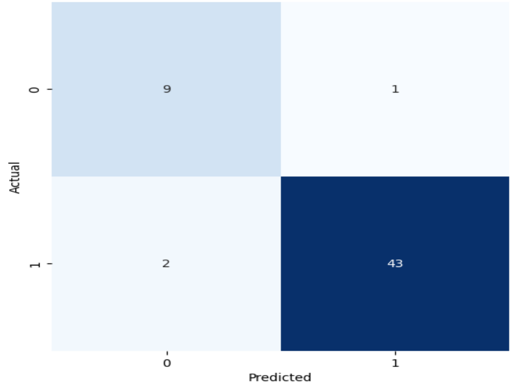

Fig 6: Confusion Matrix of the AdaBoost Model

Figure 6 displays the AdaBoost model's confusion matrix. It correctly classifies 15 actual negatives and 46 actual positives. However, 9 negatives are misclassified as positives, and 3 positives are misclassified as negatives. This suggests that the model works well, particularly when it comes to identifying the positive class.

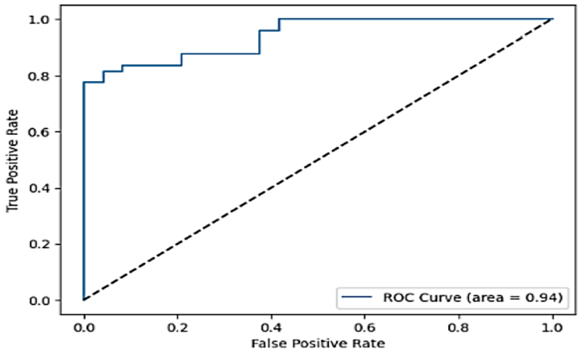

Fig 7: ROC Curve for the AdaBoost Model

Figure 7 shows the AdaBoost model's ROC curve. In terms of data classification, it does well, with an AUC of 0.94. Towards the upper left corner, the curve quickly climbs, showing high true positive rates and low false positive rates. This shows that the model is very good at telling the difference between classes.

4.1 Comparative Analysis

This section compares and contrasts the various models used for anaemia prediction. In Table 3, its present accuracy measures to see how well different models did. Anaemia can be identified using the model which shows an 89.3% accuracy, the LogitBoost [19] model. On the contrary, the accuracy of Random Forest (RF) [20] model was 67.1% which is far from desirable for Anaemia Prediction capability. AdaBoost model proved to be better than both and outperformed, with 92.7% accuracy in classification of Anaemia Prediction. However, this comparison shows that the performance of different models across this application space may differ, with AdaBoost proving to be the most effective.

Table 3. Comparative Analysis of proposed and baseline Models for Anaemia Prediction

Performance Metric	Logitboost [19]	RF[20]	AdaBoost Model
Accuracy	89.3	67.1	92.7

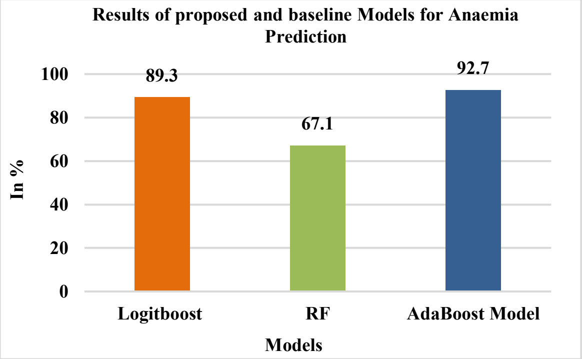

Fig 8: Results of proposed and baseline Models for Anaemia Prediction

In Figure 8 Shows the approach proposed is advantageous in anaemia detection and classification using modern ml techniques. The methodology obtains high accuracy by taking advantage of the AdaBoost model, beating previous anaemia diagnosing techniques. Through the automated feature selection process, more manual intervention is avoided and the data preprocessing is made a less daunting process. Data cleaning, one hot encoding and Z-score normalization makes data neat thus enhances data quality and the robustness of the model. A critical focus on features that improve the prediction capacity such as age, sex and hemoglobin levels, while trading off complexity of the model, results in a better model for prediction. The strong generalization ability, with little overfitting, of the model implies good promise for reliable real-time diagnostics, allowing for timely interventions to save lives improving patient outcomes in the clinical settings. Ultimately, this method improves the diagnostic ability for anemia and ultimately improves the delivery of better healthcare and the allocation of healthcare resources.

5. CONCLUSION AND FUTURE SCOPE

The accurate detection of anaemia is important in today's healthcare systems because misdiagnosis can result in severe health consequences and additional treatment costs for anaemia. ML has become a useful tool to improve diagnostic accuracy in predicting anaemia conditions given hematological data. Early detection of anaemia, early classification of anaemia is very helpful for early detection of treatment and better outcomes for the patient. Utilizing performance criteria like as recall, precision, accuracy, and F1 score, train and evaluate the AdaBoost model. It's amazing, 92.7% accurate at being able to tell an anaemia from a non-anaemia. Therefore, compare AdaBoost to other ML algorithms such as LogitBoost and Random Forest, to illustrate why it is so much better at pinpointing anaemia. It may be very good at what it does, but it will make some small mistakes and only work on a certain dataset, that is, it wouldn't work as well with large groups of people. One big problem with this study is that it only looked at data from one area, which means that the results may not apply to other places. The completeness and quality of the CBC data also affect the model's performance. In the future, researchers will look into ways to make the model more reliable and useful in real-life clinical settings. These include cross-validation with different datasets, feature engineering, and ensemble methods. Moreover, the integration of longitudinal patient data and electronic health records could enhance predictive power by capturing trends over time. Incorporating explainable AI (XAI) techniques may also improve clinician trust and adoption by making predictions more transparent.

6. REFERENCES

[1] Santorelli, A., Abbasi, B., Lyons, M., Hayat, A., Gupta, S., O'Halloran, M., & Gupta, A. (2018). Investigation of Anemia and the Dielectric Properties of Human Blood at Microwave Frequencies. IEEE Access, 6, 56885–56892. https://doi.org/10.1109/ACCESS.2018.2873447

[2] Kolluri, V. (2016b). Machine Learning in Managing Healthcare Supply Chains: How Machine Learning Optimizes Supply Chains, Ensuring the Timely Availability of Medical Supplies. Journal of Emerging Technologies and Innovative Research, 3(6).

[3] Garg, S. (2019a). AI/ML Driven Proactive Performance Monitoring, Resource Allocation and Effective Cost Management in SAAS Operations. International Journal of Core Engineering & Management, 6(6), 263–273.

[4] Kolluri, V. (2015). A Comprehensive Analysis on Explainable and Ethical Machine: Demystifying Advances in Artificial Intelligence. TIJER - International Research Journals (TIJER), 2(7).

[5] Garg, S. (2019b). Predictive Analytics and Auto Remediation using Artificial Inteligence and Machine learning in Cloud Computing Operations. International Journal of Innovative Research in Engineering & Multidisciplinary Physical Sciences, 7(2).

[6] Mutter, S., Casey, A. E., Zhen, S., Shi, Z., & Mäkinen, V. P. (2017). Multivariable Analysis of Nutritional and Socio-Economic Profiles Shows Differences in Incident Anemia for Northern and Southern Jiangsu in China. Nutrients, 9(10). https://doi.org/10.3390/nu9101153

[7] Kolluri, V. (2016a). An Innovative Study Exploring Revolutionizing Healthcare with AI: Personalized Medicine: Predictive Diagnostic Techniques and Individualized Treatment. International Journal of Emerging Technologies and Innovative Research, 3(11), 2349–5162.

[8] Bregman, D. B., Morris, D., Koch, T. A., He, A., & Goodnough, L. T. (2013). Hepcidin Levels Predict Nonresponsiveness to Oral Iron Therapy in Patients with Iron Deficiency Anemia. American Journal of Hematology, 88(2), 97–101. https://doi.org/10.1002/ajh.23354

[9] Balasubramanian, A. (2019). Intelligent Health Monitoring: Leveraging Machine Learning and Wearables for Chronic Disease Management and Prevention. International Journal of Innovative Research in Engineering & Multidisciplinary Physical Sciences, 7(6), 1–13.

[10] Esteban-Medina, M., Peña-Chilet, M., Loucera, C., & Dopazo, J. (2019). Exploring the druggable space around

the Fanconi anemia pathway using machine learning and mechanistic models. BMC Bioinformatics, 20(1), 1–15. https://doi.org/10.1186/s12859-019-2969-0

[11] Hortinela, C. C., Balbin, J. R., Fausto, J. C., Daniel C. Divina, P., & Felices, J. P. T. (2019). Identification of Abnormal Red Blood Cells and Diagnosing Specific Types of Anemia Using Image Processing and Support Vector Machine. 2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), 1–6. https://doi.org/10.1109/HNICEM48295.2019.9072904

[12] Hafeel, A., Fernando, H. S. M. H., Pravienth, M., Lokuliyana, S., Kayanthan, N., & Jayakody, A. (2019). IoT Device to Detect Anemia: A Non-Invasive Approach with Multiple Inputs. 2019 International Conference on Advancements in Computing (ICAC), 392–397. https://doi.org/10.1109/ICAC49085.2019.9103391

[13] Gebreweld, A., & Tsegaye, A. (2018). Prevalence and Factors Associated with Anemia among Pregnant Women Attending Antenatal Clinic at St. Paul's Hospital Millennium Medical College, Addis Ababa, Ethiopia. Advances in Hematology, 2018. https://doi.org/10.1155/2018/3942301

[14] Roychowdhury, S., Sun, D., Bihis, M., Ren, J., Hage, P., & Rahman, H. H. (2017). Computer Aided Detection of Anemia-Like Pallor. 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 461–464. https://doi.org/10.1109/BHI.2017.7897305

[15] Hennek, J. W., Kumar, A. A., Wiltschko, A. B., Patton, M. R., Lee, S. Y. R., Brugnara, C., Adams, R. P., & Whitesides, G. M. (2016). Diagnosis of Iron Deficiency Anemia Using Density-Based Fractionation of Red Blood Cells. Lab on a Chip, 16(20), 3929–3939. https://doi.org/10.1039/C6LC00875E

[16] Punter-Villagrasa, J., Cid, J., Páez-Avilés, C., Rodríguez-Villarreal, I., Juanola-Feliu, E., Colomer-Farrarons, J., & Miribel-Català, P. (2015). An Instantaneous Low-Cost Point-of-Care Anemia Detection Device. Sensors, 15(2), 4564–4577. https://doi.org/10.3390/s150204564

[17] Niu, B., Ren, J., & Li, X. (2019). Credit Scoring Using Machine Learning by Combing Social Network Information: Evidence from Peer-to-Peer Lending. Information, 10(12). https://doi.org/10.3390/info10120397

[18] Cao, Y., Miao, Q.-G., Liu, J.-C., & Gao, L. (2013). Advance and Prospects of AdaBoost Algorithm. Acta Automatica Sinica, 39(6), 745–758. https://doi.org/10.1016/S1874-1029(13)60052-X

[19] Dimauro, G., Guarini, A., Caivano, D., Girardi, F., Pasciolla, C., & Iacobazzi, A. (2019). Detecting Clinical Signs of Anaemia From Digital Images of the Palpebral Conjunctiva. IEEE Access, 7, 113488–113498. https://doi.org/10.1109/ACCESS.2019.2932274

[20] Anand, P., Gupta, R., & Sharma, A. (2019). Prediction of Anaemia among children using Machine Learning Algorithms. International Journal of Electronics Engineering, 11(2), 469–480.

[21] Kalla, D., Smith, N., & Samaah, F. (2023). Satellite Image Processing Using Azure Databricks and Residual Neural Network. International Journal of Advanced Trends in Computer Applications, 9(2), 48-55.

[22] Kuraku, D. S., & Kalla, D. (2023). Phishing Website URL's Detection Using NLP and Machine Learning Techniques. Journal on Artificial Intelligence-Tech Science.

[23] Varadharajan, V., Smith, N., Kalla, D., Samaah, F., Polimetla, K., & Kumar, G. R. (2024). Stock Closing Price and Trend Prediction with LSTM-RNN. Journal of Artificial Intelligence and Big Data, 4, 877.

[24] Kalla, D., Kuraku, D. S., & Samaah, F. (2021). Enhancing cyber security by predicting malwares using supervised machine learning models. International Journal of Computing and Artificial Intelligence, 2(2), 55-62.

[25] Kuraku, D. S., Kalla, D., Smith, N., & Samaah, F. (2023). Safeguarding FinTech: elevating employee cybersecurity awareness in financial sector. International Journal of Applied Information Systems (IJAIS), 12(42).

[26] Kalla, D., Smith, N., Samaah, F., & Polimetla, K. (2022). Enhancing Early Diagnosis: Machine Learning Applications in Diabetes Prediction. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-205. DOI: doi. org/10.47363/JAICC/2022 (1), 191, 2-7.

[27] Kuraku, S., Kalla, D., Samaah, F., & Smith, N. (2023). Cultivating proactive cybersecurity culture among IT professional to combat evolving threats. International Journal of Electrical, Electronics and Computers, 8(6).

[28] Kalla, D., Smith, N., Samaah, F., & Polimetla, K. (2024). Hybrid Scalable Researcher Recommendation System Using Azure Data Lake Analytics. Journal of Data Analysis and Information Processing, 12, 76-88.

[29] Kuraku, D. S., & Kalla, D. (2023). Impact of phishing on users with different online browsing hours and spending habits. International Journal of Advanced Research in Computer and Communication Engineering, 12(10).

[30] Kalla, D., & Kuraku, S. (2023). Phishing website url's detection using nlp and machine learning techniques. Journal of Artificial Intelligence, 5, 145.

[31] Kuraku, D. S., Kalla, D., & Samaah, F. (2022). Navigating the link between internet user attitudes and cybersecurity awareness in the era of phishing challenges. International Advanced Research Journal in Science, Engineering and Technology, 9(12).

[32] Kuraku, D. S., Kalla, D., Smith, N., & Samaah, F. (2023). Exploring How User Behavior Shapes Cybersecurity Awareness in the Face of Phishing Attacks. International Journal of Computer Trends and Technology.

[33] Sreeramulu, M. D., Mohammed, A. S., Kalla, D., Boddapati, N., & Natarajan, Y. (2024, September). AI-driven Dynamic Workload Balancing for Real-time Applications on Cloud Infrastructure. In 2024 7th International Conference on Contemporary Computing and Informatics (IC3I) (Vol. 7, pp. 1660-1665). IEEE.

[34] Kalla, D., Mohammed, A. S., Boddapati, V. N., Jiwani, N., & Kiruthiga, T. (2024, November). Investigating the

Impact of Heuristic Algorithms on Cyberthreat Detection. In 2024 2nd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT) (Vol. 1, pp. 450-455). IEEE.

[35] Chandrasekaran, A., & Kalla, D. (2023). Heart disease prediction using chi-square test and linear regression. Computer Science & Information Technology, 13, 135-146.

[36] Chinta, P. C. R., Katnapally, N., Ja, K., Bodepudi, V., Babu, S., & Boppana, M. S. (2022). Exploring the role of neural networks in big data-driven ERP systems for proactive cybersecurity management. Kurdish Studies.

[37] Routhu, K., Bodepudi, V., Jha, K. M., & Chinta, P. C. R. (2020). A Deep Learning Architectures for Enhancing Cyber Security Protocols in Big Data Integrated ERP Systems. Available at SSRN 5102662.

[38] Moore, C. (2023). AI-powered big data and ERP systems for autonomous detection of cybersecurity vulnerabilities. Nanotechnology Perceptions, 19, 46-64.

[39] Bodepudi, V., & Chinta, P. C. R. (2024). Enhancing Financial Predictions Based on Bitcoin Prices using Big Data and Deep Learning Approach. Available at SSRN 5112132.

[40] Chinta, P. C. R. (2023). The Art of Business Analysis in Information Management Projects: Best Practices and Insights.

[41] Boppana, S. B., Moore, C. S., Bodepudi, V., Jha, K. M., Maka, S. R., & Sadaram, G. AI And ML Applications In Big Data Analytics: Transforming ERP Security Models For Modern Enterprises.

[42] Katnapally, N., Chinta, P. C. R., Routhu, K. K., Velaga, V., Bodepudi, V., & Karaka, L. M. (2021). Leveraging Big Data Analytics and Machine Learning Techniques for Sentiment Analysis of Amazon Product Reviews in Business Insights. American Journal of Computing and Engineering, 4(2), 35-51.

[43] Chinta, P. C. R., Moore, C. S., Karaka, L. M., Sakuru, M., Bodepudi, V., & Maka, S. R. (2025). Building an Intelligent Phishing Email Detection System Using Machine Learning and Feature Engineering. European Journal of Applied Science, Engineering and Technology, 3(2), 41-54.

[44] Moore, C. (2024). Enhancing Network Security With Artificial Intelligence Based Traffic Anomaly Detection In Big Data Systems. Available at SSRN 5103209.

[45] Chinta, P. C. R., Moore, C. S., Karaka, L. M., Sakuru, M., & Bodepudi, V. (2025). Predictive Analytics for Disease Diagnosis: A Study on Healthcare Data with Machine Learning Algorithms and Big Data. J Cancer Sci, 10(1), 1.

[46] KishanKumar Routhu, A. D. P. Risk Management in Enterprise Merger and Acquisition (M&A): A Review of Approaches and Best Practices.

[47] Bodepudi, V. (2023). Understanding the Fundamentals of Digital Transformation in Financial Services: Drivers and Strategic Insights. Journal of Artificial Intelligence and Big Data, 3(1), 10-31586.

[48] Chinta, P. C. R. (2022). Enhancing Supply Chain Efficiency and Performance Through ERP Optimisation Strategies. Journal of Artificial Intelligence & Cloud Computing, 1(4), 10-47363.

[49] Krishna Madhav, J., Varun, B., Niharika, K., Srinivasa Rao, M., & Laxmana Murthy, K. (2023). Optimising Sales Forecasts in ERP Systems Using Machine Learning and Predictive Analytics. J Contemp Edu Theo Artific Intel: JCETAI-104.

[50] Jha, K. M., Velaga, V., Routhu, K., Sadaram, G., Boppana, S. B., & Katnapally, N. (2025). Transforming Supply Chain Performance Based on Electronic Data Interchange (EDI) Integration: A Detailed Analysis. European Journal of Applied Science, Engineering and Technology, 3(2), 25-40.

[51] Sadaram, G., Sakuru, M., Karaka, L. M., Reddy, M. S., Bodepudi, V., Boppana, S. B., & Maka, S. R. (2022). Internet of Things (IoT) Cybersecurity Enhancement through Artificial Intelligence: A Study on Intrusion Detection Systems. Universal Library of Engineering Technology, (2022).

[52] Maka, S. R. (2023). Understanding the Fundamentals of Digital Transformation in Financial Services: Drivers and Strategic Insights. Available at SSRN 5116707.