Automatika

Journal for Control, Measurement, Electronics, Computing and Communications

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/taut20

Anemia detection and classification from blood samples using data analysis and deep learning*

Nilesh Bhaskarrao Bahadure, Ramdas Khomane & Aditya Nittala

To cite this article: Nilesh Bhaskarrao Bahadure, Ramdas Khomane & Aditya Nittala (2024) Anemia detection and classification from blood samples using data analysis and deep learning*, Automatika, 65:3, 1163-1176, DOI: 10.1080/00051144.2024.2352317

To link to this article: https://doi.org/10.1080/00051144.2024.2352317

Published online: 14 May 2024.

AUTOMATIKA
2024, VOL. 65, NO. 3, 1163–1176
https://doi.org/10.1080/00051144.2024.2352317

Anemia detection and classification from blood samples using data analysis and deep learning*

Nilesh Bhaskarrao Bahadure, Ramdas Khomane and Aditya Nittala

Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, India

ABSTRACT

This study aims to examine the possibility and impact of utilizing data science on blood samples to rapidly and proactively identify underlying health issues. By utilizing effective algorithms, models will be constructed to address these problems and determine potential healthcare options based on geographical location. Once data is gathered, health officials will be notified of major diseases and individuals at risk or already affected. Authentic blood samples are used to ensure the credibility and validity of the proposed system. The data was collected during a volunteer-led hemoglobin blood test camp specifically for women residing in impoverished areas, resulting in a total of 551 samples. The effectiveness of this technique has been assessed through experimental results based on Hb, RDW%, MCV, RBC, and M-Index. The proposed data analysis and deep learning algorithm achieved average values of haemoglobin count 11.67 g/dL with a 1.33 standard deviation, RDW 14.59%, MCV 81.45, RBC 4.37 per microliter with a variance of 0.5, and M-Index 19.56. The experimental results achieved 97.60% accuracy, demonstrating the effectiveness of the proposed technique for classifying anemia and its subtypes. The experimental results indicate better overlap between the automated identification of anemia and manual detection by the experts.

ARTICLE HISTORY
Received 24 November 2023
Accepted 2 May 2024

KEYWORDS
Deep learning; hemoglobin (Hb); Red Blood Cell Distribution Width (RDW); mean corpuscular volume (MCV); Red Blood Cell count (RBC); Mentzer Index (M-Index)

1. Introduction

In healthcare, the combination of medical diagnostics and cutting-edge technology has revolutionized how we detect and improve patient outcomes. One research area of study focuses on using deep learning techniques to analyze blood samples to detect Anemia, a commonly misdiagnosed blood disorder [1,2]. Timely detection and appropriate treatment are crucial steps in curing Anemia, as it can lead to significant health risks if left untreated. Symptoms like fatigue, weakness, and shortness of breath are critical indicators of Anemia, typically caused by low hemoglobin levels or a deficiency in red blood cells. Traditionally, hematologists or clinical experts would manually examine blood samples for Anemia, and the process was time-consuming and prone to human error [3,4]. However, recent developments in Information Technology and e-healthcare systems, driven by artificial intelligence and deep learning, have transformed the field of medical diagnostics [5–9].

With the arrival of this new era, a large amount of collection of blood samples can be utilized to train algorithms that rely on neural networks or other systems rooted in learning. The assessment of these blood samples can now be completed in a fraction of the time. This innovative technology, particularly pattern recognition, has the potential to significantly enhance the efficiency and precision of detecting and diagnosing anemia [10,11].

This study attempts a serious dominion of data analysis of blood samples using advanced methodologies based on deep learning with a distinct focus on identifying Anemia. The blood samples consist of a vast repository of information; by exploiting that information, we look to develop a robust mechanism capable of not only detecting and identifying the presence of Anemia but also sensing valuable information on different subtypes and severity levels [12–15].

The analysis process is fully automated, and the blood samples are not taken from pre-existing datasets. Instead, they are collected near Nagpur city in Maharashtra state, India. In Nagpur, a non-governmental organization (NGO) called Youth for Seva. Its primary objective is to motivate and engage young individuals in volunteering activities, providing them with meaningful opportunities to contribute to their community. With the assistance of a local hospital, Youth for Seva organized a blood donation camp in an underprivileged area near Shukrawari Lake in Nagpur. The reason for selecting the Shukrawari Lake area in Nagpur is, this locality has many underprivileged communities, and our work primarily focuses on detecting and classifying

CONTACT Nilesh Bhaskarrao Bahadure nilesh.bahadure@sitnagpur.siu.edu.in

*All the authors contributed to the conception and design of the study. Data were collected by NBB, and AN. NBB and RK performed the data analysis. The first draft of the manuscript was written by NBB and AN, and RK reviewed and edited the manuscript. All authors read and approved the final manuscript.

anemia and its subtypes in the underprivileged area. A total of 551 samples were collected: 384 (70%) were used for training, and the remaining were used for testing the algorithm. These samples are enough for training and testing the algorithm and validating the performance of the proposed system. This initiative allows government authorities to gain insight into the health situation in that specific vicinity, enabling them to take appropriate action as necessary.

Typically, laboratories are supervised by clinical experts or hematologists; they collect the donor's Age and Sex information. They extract essential attributes of the blood such as hemoglobin count, Red Blood Cell count, Red Cell Distribution Width percentage, Mean Corpuscular Volume, and Mentzer index from the provided blood samples.

Gathering this data, understanding how these constituents of blood, and applying data science to it will reduce the time health professionals need to take appropriate actions in providing adequate care under government schemes.

Constituents of blood-related to different illnesses and identification of localities, comprising of people with different immunities, results in a machine learning model which can target a locality in a time of need [16]. An NGO can set up a blood donation camp, which can quickly fulfil the requirement.

The entire process is highly straightforward and streamlines the diagnostic process with deep learning, thus enabling swift and precise identification of Anemia in a diverse patient population. Furthermore, the self-operating nature of deep learning models brings down the dependence on manual processing by clinical experts or hematologists and thus improves the accuracy of the Anemia diagnosis and is within the time frame, and thus offers significant aid to the healthcare [17–19].

The research also aims to develop an accurate anemia and subtype detection system. It utilizes deep learning-based YOLO techniques, specifically a deep learning-based YOLO model for detection and classification. The ultimate objective is to improve the classification accuracy of anemia and subtype detection, thereby enabling early prediction and potentially saving human lives. This paper addresses one of the most underprivileged areas in Nagpur, India, locality, and its findings can encourage the development of a similar solution to other geographic locations or populations.

The core objective of this study is to integrate the power of automation through advanced technology such as deep learning models, fine-tuning them to handle the variation of blood sample data, identifying the severity levels for Anemia and other subtypes of diseases, and validating the efficacy of our proposed system through rigorous testing and analysis [20,21]. The ultimate aim is to devise a solution for one of the life-threatening diseases using the intersection of healthcare and artificial intelligence, with a substantial impact on the early detection, classification, and identification of Anemia.

2. Related works

Anemia is a severe disease that poses a threat to life and contributes to the development of many other diseases. In addition to Anemia, individuals with this illness may also be susceptible to other blood-related diseases. To successfully address this issue, it is crucial to identify the disease in its early stages, helping healthcare professionals to recommend appropriate treatments. However, a traditional manual detection method is time-consuming and inaccurate. Fortunately, numerous researchers have offered advanced technologies that streamline blood sample collection and enable timely and precise disease detection.

Zhao et al. [19] investigated the machine learning-based system for predicting COVID-19 disease and Pneumonia patients. Different machine learning algorithms such as Random Forest, XGBoost, Logistic regression, and deep learning neural networks were employed for the severity prediction. Detailed statistical analysis and feature interpretation were utilized to understand the relationship between clinical variables and disease outcomes. The results in terms of clinical indicators are promising and may help doctors predict the progression and spread of COVID-19 and other types of pneumonia.

A simple paper-based technique measuring blood hemoglobin is investigated by Xiaoxi Yang et al. [22]. A 20 μL droplet of a mixture of blood and Drabkin reagent was deposited, and the resulting bloodstain was digitized for further analysis. The bloodstain colour intensity was used to measure Hb. A total of 54 subjects were analyzed, and the performance of the paper-based Hb with a hematology analyzer was compared. Detection of DNA damage from the blood samples was proposed by Kristina Schulze Johann et al. [23]. They calculated the degradation index (DI) to assess DNA quality.

Vijayarani and Sudha [24] performed the prediction of various diseases from hemogram blood samples using data mining techniques. They employed a weight-based K- K-means algorithm for classifying leukemia, bacterial, HIV infection, inflammatory, and pernicious anemia. The detailed analysis of Fuzzy C-means and K-means clustering with the proposed weight-based K-means algorithm is also evaluated. Eric Boersma et al. [25] suggested a study to identify the acute risk of coronary syndrome. The detailed evolution of blood biomarkers has been studied for patients with post-acute coronary syndrome. A total of 844 patients' data is used for the analysis, and clinical observations have been performed over a certain period.

The base of this entire study started after the collection of blood samples. The main problem is identifying suitable and willing donors and collecting the blood samples. A blood transfusion is performed to ensure the blood is free from any infection. Ahdan and Setiawansyah [26] devise a solution to this problem. The advancement of information technology is integrated with the geolocation system and use the Dijkstra algorithm to find the closest route for finding blood banks and donors. The system can identify and recommend the donors according to the patients' qualifications (specific requirements). Hai Trieu Le et al. [27] also investigated the same BloodChain system using blockchain technology.

Lamia Alhazmi [28] introduces a novel method exploiting a deep learning approach to identify white blood cells (WBC), red blood cells (RBC), and platelets in blood samples. In the diagnosis process of most diseases, the initial step is a blood test, allowing for the evaluation of various quality measures for finding the root cause of the disease. However, manual detection may lead to inaccuracies and be time-consuming. The proposed study offers a significant integration of technology that enables the automatic detection of cells in blood samples. Through specific proposed training of the model, they achieved reliable results, with 100% accuracy in counting WBC, 89% in RBC, and 96% in platelets.

A detailed literature review has been done, and its core finding is summarized in Table 1.

3. Materials and methods

This section presents the materials, the source of the blood samples, and the algorithm used to detect and classify Anemia. Figure 1 provides the flow diagram of the algorithm. Data was collected during a volunteer-run hemoglobin blood test camp for women living in impoverished areas, organized in cooperation with a nearby non-governmental organization (NGO). Five hundred fifty-one women and 130 men took part in the campaign. Figure 2 shows some of the sample data used for testing the algorithm. The gathered data was then digitalized to be analyzed.

Several quality matrices become significant for assessing and interpreting the blood sample data for automatically detecting Anemia and its subtypes through deep learning techniques. How the entire dataset is distributed, its training model and its characteristics are easily understood using these quality matrices. They are a great help in the identification of probable anomalies or motifs correlated with Anemia and its subtype. The details about the quality matrices used in this study are discussed below.

(1) Minimum value (min): This represents the lowest value in all the entity contexts used in the Anemia detection, such as age, hemoglobin count, red blood cells, RDW%, MCV, and M-Index. It evidences a crucial indicator and signifies the lowest point in the dataset distribution.
(2) Maximum value (max): This represents the highest value in all the entity contexts used in Anemia detection, such as age, hemoglobin count, red blood cells, RDW%, MCV, and M-Index. It evidences a crucial indicator and signifies the highest point in the dataset distribution. It can also assist in identifying extreme or uncommonly high readings that might advocate other severe health issues or motifs in the data.
(3) Average (mean): The central tendency and pattern of the data distribution in the dataset is provided employing average value. In the case of Anemia and its subtype detection, this could indicate and signify the average count of age, hemoglobin count, red blood cells, RDW%, MCV, and M-Index, thus giving a general idea of the "distinctive" value in the dataset.
(4) Mode: Some of the values occurred frequently in the dataset; the mode is the value that signifies this repetitive pattern. In the context of the diagnosis of Anemia, how often the red blood cell count or hemoglobin level occurs is very important, and the mode is indicated the same for the dataset used.
(5) Variance: The spread or dispersion of the data around the mean value is indicated by variance. In the case of Anemia and its subtype detection, the higher variance suggests the wide spread of the data points, and the low indicates clustered around the mean. Variance may suggest probable irregularities in the entity, such as age, hemoglobin count, red blood cells, RDW%, MCV, and M-Index across the dataset.
(6) Standard deviation: The average amount of variation or dispersion of the data from the mean is indicated by the standard deviation parameter. In the case of Anemia and its subtype detection, the higher standard deviation may suggest a broader range of values, and the low suggests close to the mean, indicating probable irregularities in the entity such as age, hemoglobin count, red blood cells, RDW%, MCV, and M-Index across the dataset.

Any irregularities in the entity are easily accessed and identified through the quality matrices and could signal different stages, types, or levels of anemia. Moreover, understanding these quality matrices can assist in setting relevant thresholds in the diagnosis of anemia and also help to keep an eye on the efficacy of the model's predictions.

The experimentation is performed using an Apple MacBook M2 system equipped with 16 gigabytes of RAM and 256 gigabytes of storage. The software

Table 1. Summary of literature review.

Authors	Year	Techniques Used	Findings
Dimauro et al. [29]	2023	Deep Learning with RUSBoost	The RUSBoost technique is used for anemia detection on a joint dataset of Italian and Indian patient data. Holdout cross-validation (CV) was employed in the experimentation with 200 iterations and achieved an accuracy of 83%. This study counsels that class imbalance has a jolt on the automated detection of anemia.
Haggenmuller et al. [30]	2023	Mobile App	This study presents the self-monitoring of anemia detection through the mobile App, but its use is limited to close monitoring of Hb concentration only.
Dimauro et al. [31]	2023	Sclera segmentation algorithm	The sclera and scleral blood vessels from the eye images were extracted, and anemia detection was performed by analyzing the region of interest. The region of interest is segmented using the sclera algorithm, extracted using a vessel algorithm, and then using the classifi er; fi nally, anemic status or normal controls are predicted. The results were promising, with precision 88.53, F1 84.10, and recall 82.53.
Dhalla et al. [32]	2023	Deep neural algorithm	The ve pre-trained segmentation deep learning-based architectures, namely UNet, UNet++, FCN, PSPNet, and LinkNet, are used for conjunctiva segmentation for anemia detection. The LinkNet architecture outperforms with an accuracy of 94.17% compared to its counterparts. This study proves the signi ficance of LinkNet architecture for real-time segmentation of palpebral conjunctiva for anemia detection from eye-contaminated images.
Saputra et al. [33]	2023	Extreme learning machine	An arti ficial intelligence-based technique using an extreme learning machine is used to predict and detect anemia. This technique achieved an accuracy of 99.21%. The authenticity of the proposed system is questionable as it used only 63 test images.
Kistenev et al. [34]	2022	Machine learning	The use of standard blood testing and machine learning for COVID-19 detection is addressed. To increase accuracy and efficiency of existing diagnostic techniques, such as the PCR test, and recommends integrating clinical tests with artifi cial intelligence.
Kukar et al. [35]	2021	Extreme Gradient	The model was developed using information from 160 patients who tested positive for COVID-19 and 5333 patients who had viral and bacterial illnesses.
Chen et al. [36]	2020	Mathematical statistics analysis	In the presented study, the causes of the poor quality of blood samples used in medical examinations are covered. Haemolysis, coagulation, partial anticoagulation, incorrect blood collection test tubes, inadequate preparation, insufficient specimen volume, postponed examination submission, and other factors are among the primary factors that were found.
Pfeil et al. [37]	2019	Machine learning and Deep learning	The outcomes show that blood disorders can be successfully predicted using classical machine learning algorithms using standard blood tests.
Alsheref and Gomaa [38]	2019	Machine learning and Deep learning	The outcomes show that blood disorders can be successfully predicted using classical machine learning algorithms using standard blood tests.
Noor et al. [39]	2019	Data pre-processing and different Machine learning algorithms	This research focuses on an efficient approach for applying machine learning algorithms to determine hemoglobin (Hb) levels. The typical methods for determining hemoglobin levels are expensive, time-consuming, and intrusive.
Golap et. al. [40]	2019	Multi-Gene Genetic Programming-Based model and Machine learning	In this study, 39 time-domain and 6 frequency-domain variables that were obtained from PPG signals were attached with gender and age. To choose the optimal characteristics for training and creating a mathematical model, a correlation-based feature selection method was used.
Gincar et. al. [41]	2018	Machine learning algorithms (CART decision tree and Random forest)	Based on the findings of laboratory blood tests, two predictive models for haematologic illnesses were constructed using machine learning techniques. One model made use of every blood test parameter that was available, while the other made use of a smaller dataset that was usually measured at patient arrival.

environment utilized Python 3 and Microsoft Excel 2019 for data processing and analysis.

The detailed experimentation operation with the different quality parameters is shown in Figure 3 followed by its summarization.

(1) Pre-processing Data
   (a) Convert the collected data from Excel to CSV containing blood constituents.
   (b) Format the data and handle the missing values.
(2) Data Analysis of the collected data
   (a) Calculate statistical metrics for each blood constituent in the dataset, such as mean, median, and standard deviation.
   (b) Define acceptable thresholds (lower and upper) for each constituent, which can be based on medical standards or local population averages.
(3) Identify Locality-Wide Deviations
   (a) Compare the calculated statistics with the defined thresholds.
   (b) Identify constituents that deviate significantly from the expected range for the entire locality.
   (c) Generate reports or alerts for constituents with notable deviations.
(4) Data Analysis for Individuals
   (a) For each individual in the dataset, calculate their blood constituent values.

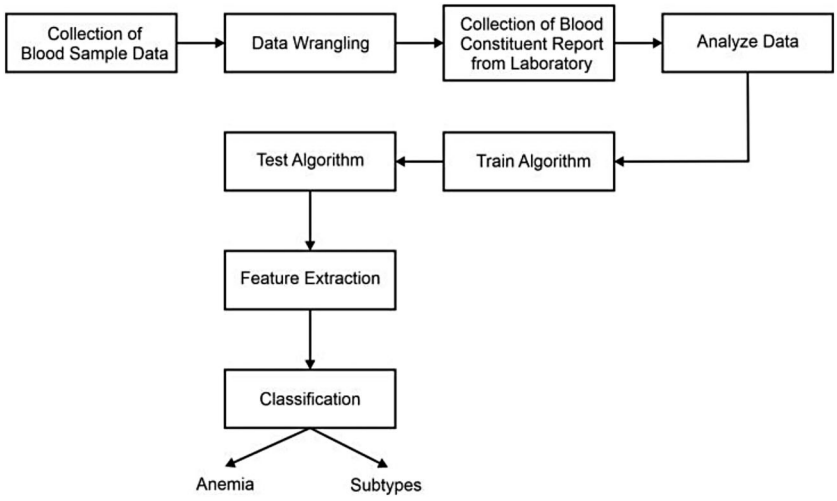

Figure 1. Steps used in proposed algorithm.

	A	C	D	E	F	G	H	I	J
1	Serial Number	Age	Sex	HB	RDW%	MCV	RBC	M.INDEX	FOR-HPLC
2	1	45	Female	12.1	13.2	80	4.68	17.09	NO
3	2	48	Female	9.8	14.5	78	3.83	20.37	NO
4	3	70	Female	12.8	15.6	70	5.74	12.2	YES
5	4	17	Female	11.2	14.8	83	4.19	19.81	NO
6	5	60	Female	11.5	14.6	82	4.27	19.2	NO
7	6	43	Female	12	14.1	85	4.32	19.68	NO
8	7	53	Female	11.5	13.1	91	3.72	24.46	NO
9	8	54	Female	9.2	20	106	2.66	39.85	NO
10	9	53	Female	12.1	14.9	75	5.02	14.94	NO
11	10	57	Female	13.2	15.4	94	4.29	21.91	NO
12	11	13	Female	12.3	14.7	95	3.88	24.48	NO
13	12	32	Female	11.3	14.6	75	4.67	16.06	NO
14	13	50	Female	12.7	13.9	78	4.93	15.82	NO
15	14	67	Female	12.5	15.1	2	4.95	15.96	NO
16	15	45	Female	12.4	13.6	81	4.78	16.95	NO

Figure 2. Sample data.

(b) Compare individual values with the established thresholds.

(5) Identify Individual Deviations
   (a) For each individual, identify constituents with values outside the expected range.
   (b) Generate individual health reports highlighting deviations and indicating potential health issues.

(6) Visualization
   (a) Create visualizations to represent data for easier understanding.

(7) Output and reporting
   (a) Generate reports for the entire locality's health profile, including identified issues.
   (b) Generate individual health reports for each person in the dataset.

(8) Alerting System (optional)
   (a) Implement an alerting system to notify healthcare providers or individuals when significant deviations are detected.

(9) Regular Updates
   (a) Set up regular data updates to ensure the analysis is based on the most recent data.

3.1. Classification

Classification is executed to extract vital information and findings from medical images. The classification achieves higher accuracy and gives valued information about the affected area by the diseases [42]. The classification complexity reduction and improvement in accuracy are noticed with the help of proper acquisition, feature extraction (extraction of quality matrices), and feature optimization (unnecessary in case only a few features or quality matrices are extracted). The suggested classification process is shown in Figure 4.

Popular classification techniques such as support vector machine (SVM), random forest (RF), Self-organizing map (SOM), and principal component analysis (PCA) classifiers are unable to support low-resolution images. Earlier work has shown that these classifiers are computationally complex and require significant time for convergence when working on larger datasets. These limitations are resolved by using the YOLOv6-based classification method [43], which is proposed in this paper.

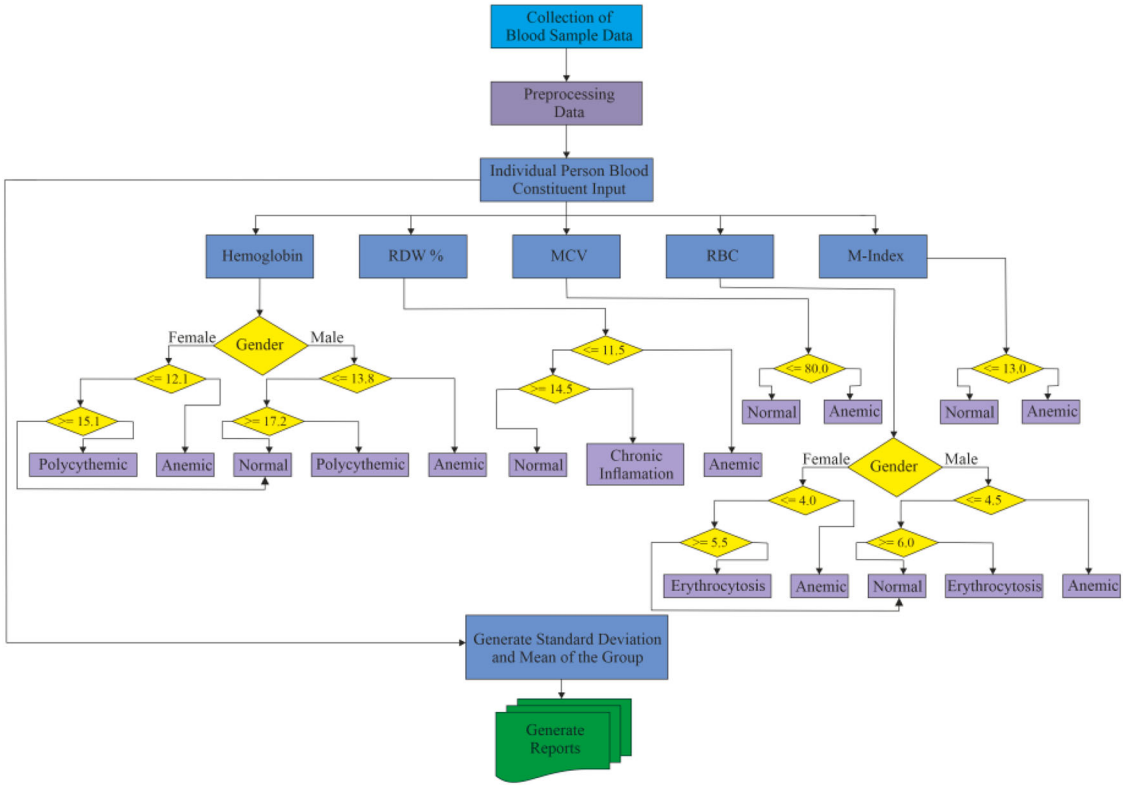

Figure 3. Experimentation process.

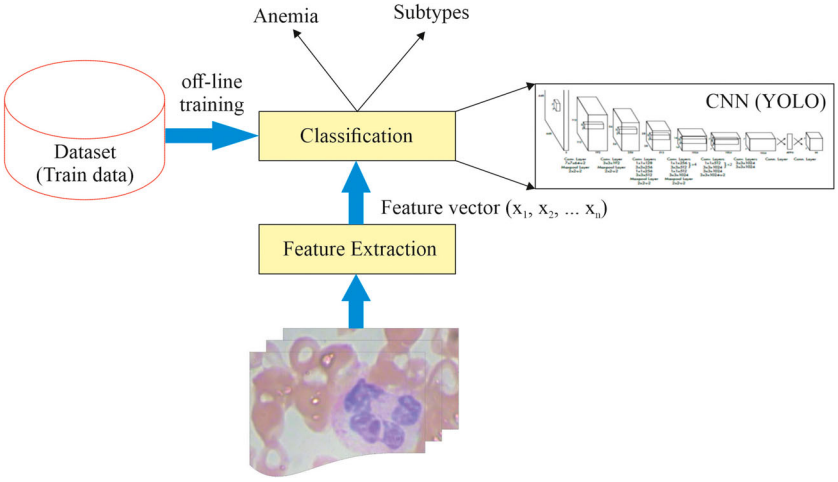

Figure 4. Process of classification.

Machine learning and deep learning techniques have gained significant popularity for tumour classification. Most of these methods involve the first step of learning from training models developed from annotated images of large datasets, where they learn about features and patterns of infected tissues. CNN-based architectures such as YOLO and SSD (Single Shot Detector) have shown promising results in brain tumour detection.

Computer vision applications use the well-known object detection algorithm YOLO. It is renowned for its real-time performance and speed. YOLO breaks up an input image into a grid of cells, using which multiple bounding boxes and class probabilities for the objects in each cell are predicted. Steps to detect objects using YOLO:

(1) Obtain a blob from the image since we require fixed-size input.
(2) Store the various layers extracted using YOLO in a variable.
(3) Forward the variable to the YOLO network and then receive the output.
(4) Store the output in the layer output variable.

The dataset is trained for 160 Epochs with the input image size 224 × 224 and 0.1 as the initial learning rate for the training purpose. During the training

process, standard data increment methods are used. Then, the fine-tuning of the network is considered using a 448 × 448 image size with the initial learning rate changed to 0.001 for 30 epochs, and the training is performed ten times.

The detection and identification often require fine-grained visual information; for this purpose, the network's input resolution has been increased from 224 × 224 to 448 × 448. The model YOLOv6 adjusts its filter to perform better on higher-resolution images, so it uses a higher resolution 448 × 448 input instead of 224 × 224. The accuracy on 224 × 224 images was calculated and achieved 93.41% (156 samples out of 167 samples detected correctly) and was taken at 6.26 s, whereas the same images on 448 × 448 resolution gave 97.60% (163 samples out of 167 samples detected correctly) accuracy and were taken at 3.05 s. This approach increased the accuracy by 4% after training for 30 epochs.

Our final layer effectively forecasts both class probabilities and bounding box coordinates. A linear activation function is employed for the final layer, and leaky rectified linear activation shown in (1) is used for all other layers.

$$\emptyset(x) = \begin{cases} x, & \text{if } x > 0 \\ 0.1x, & \text{otherwise} \end{cases} \tag{1}$$

The YOLO algorithms are strong enough to handle multi-class classification. Image or object detection consists of two tasks:

(1) Image classification
(2) Object localization

Through the image classification algorithms, the type or class of an object is predicted. In contrast, object localization algorithms find the object in the image and represent it with a bounding box.

YOLO uses one of the best architectures of neural networks. Due to its simplicity, high accuracy, and high processing speed, YOLO has become a highly preferred object detection model. It predicts a class and the bounding box that defines the object's location on the input image. Each bounding box recognizes four members:

(1) (bx; by) as the centre of the bounding box
(2) (bw) as the width of the box
(3) (b_h) as the box height

In addition to this, it predict the corresponding number c for the predicted class and probability of the prediction (Pc).

For example, the image is divided into a grid, a 3 × 3 grid. Through the grid, it becomes easy to detect one object per grid cell compared to one object per image.

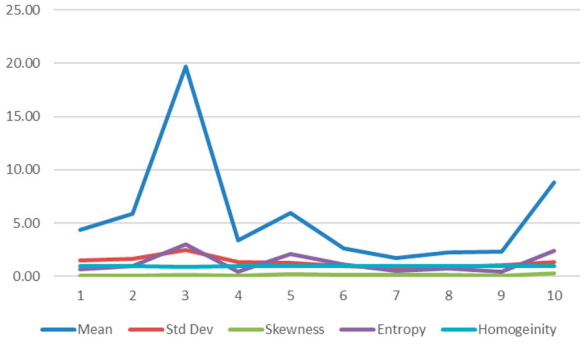

Figure 5. Plot of some prominent features.

In the next step, each grid cell is described by a vector. For example, in the case of brain MRI image, two classes are defined such as Normal and Abnormal, then it is described as:

$$C_{r,c} = (P_c, b_x, b_y, b_w, b_h, C_1, C_2)$$

Where $C_{r,c}$ represents the corresponding grid cell, for example, the first cell from the 3 × 3 grid is represented as $C_{1,1}$. P_c is the probability of the object class, b_x and by are the coordinates of the centre of the bounding box, b_h, and b_w are the height and width of the bounding box relative to the entire image, and C_1 and C_2 are represented for the class, i.e. C_1 for the "Anemia" and C_2 for the "Subtypes". The value of C_1 and C_2 is 0 and 1, depending on which class represents the bounding box. Algorithm (1) enlists various steps involved in implementing YOLOv6 for detecting and classifying brain tumours.

The YOLOv6 algorithm quickly accesses the trained data, even though it is also available in pretrained format. In our research, the classification is done to find two significant constraints: anemia and its subtypes. The classification algorithm performs the classification quickly and effectively if only relevant features are extracted and fed to the classifier for data analysis. Feature extraction brings more preciseness and clarity to the image, which defines the body's colour, texture, size, and edges. Feature extraction is crucial for reducing the classifier's complexity to classify an image's characteristics. Nearly 13 features were extracted from the blood samples. Through extensive analysis, it was found and observed that only five features are relevant and sufficient for detecting anemia and its subtypes. Table 2 gives some valuable features required to analyze the blood samples, such as mean, standard deviation, Skewness, Kurtosis, and Homogeneity. Figure 5 plots these prominent features for ten randomly selected images.

4. Results and discussion

As we age, a small change occurs in the blood's chemical composition. This aging process leads to a decrease in the total body water, resulting in a drop in the fluid

Table 2. Prominent features.

Metrics	Formulae
Mean (M)	$\mathrm{M} = \left(\frac{1}{n\times m}\right) \sum_{0}^{n-1} \sum_{0}^{m-1} f(n,m)$ where n and m are image size. A lower value indicates good amount of noise elimination from the image.
Standard deviation (SD)	$SD = \sqrt{\left(\frac{1}{n\times m} \sum_{0}^{n-1} \sum_{0}^{m-1} (f(n,m) - M)^2\right)}$ A higher value indicates better intensity level and high contrast among edges of an image.
Entropy (E)	$E = -\sum_{m=0}^{m-1} \sum_{n=0}^{n-1} f(n,m) log2 f(n,m)$ Higher value of entropy indicates more information contents and also indicates better imperceptibility.
Skewness (Sk)	$\mathrm{S_k} = \frac{1}{m\times n} \frac{\sum \lvert (f(n,m)-M)^3 \rvert}{SD^3}$ Skewness is a measure of symmetry or the lack of symmetry. Low value indicates better anemia detection.
Homogeneity (H)	$H = \sum_{m=0}^{m-1} \sum_{n=0}^{n-1} \frac{1}{1+(x-y)^2} f(x,y)$ Homogeneity may have a single or a range of values so as to determine whether the image is textured or non-textured. Less or no variation indicates better anemia detection.

content and blood volume. Additionally, evidence suggests a decrease in the production of red blood cells (RBCs) in response to disease or stress. A total of 551 female participants provided samples, ranging in age from 13 to 82 years. It was observed that the average age of the women at the camp was 49 years old. Most of the women in the camp have a mean age of 60 years, clustered around this age with a standard deviation of 17 years.

Data regarding age provides valuable insights into the socioeconomic status and demographic characteristics of the women residing near Juni Shukravari. It is worth mentioning that these women belong to the lower-income bracket and are facing economic hardships.

Algorithm 1 Classification algorithm using YOLO

1. Import the required packages, and libraries
2. Select threshold value (0.5), box condence score, and box class probability
3. Calculate score, boxes, and classes
4. Calculate IoU between two boxes
$$\mathrm{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$
5. Select non-max suppression
6. Select value of the shape (19, 19, 5, 7) randomly and then predict the bounding boxes
$$\mathbf{Y} = \begin{matrix} P_c \\ b_x \\ b_y \\ b_h \\ b_w \\ C_1 \\ C_2 \end{matrix}$$
7. Generate suppressed boxes from the output of CNN
8. Find the prediction for a random volume
9. Apply pre-trained YOLO algorithm on new images
10. Generate the prediction of bounding boxes and save the images (Im1)
11. Get an image and make predictions using the predict function
12. Plot the predictions

A measure used to measure the amount of hemoglobin in the bloodstream is called hemoglobin concentration (Hb). The reference range for hemoglobin levels in adults is defined as follows: Adult males: hemoglobin levels ranging from 13.8 to 17.2 grams per decilitre (g/dL).

Adult females: hemoglobin levels ranging from 12.1 to 15.1 g/dL.

Figure 6 shows the scatter plot between the age of the participants and their Hb count value. The investigation showed an average hemoglobin count of 11.67 g/dL, with a 1.33 standard deviation.

None of the women assessed had hemoglobin levels higher than 15 g/dL; the maximum amount measured was 14.2 g/dL (minimum: 6.9 g/dL). These results show that women in this location are more vulnerable to anemia, but they are less likely to be polycythemia-prone.

RDW (Red Blood Cell Distribution Width) is a measure that shows how variable the size of the red blood cells is inside a blood sample. Increased RDW% levels indicate increased variability in red blood cell sizes, which is linked to the emergence of several medical disorders, including different types of anemia. On the other hand, lower RDW% values are associated with particular diseases marked by an abnormal consistency in red blood cell size.

Figure 7 shows the scatter plot between the age of the participants and their RDW% value. The analysis showed that the RDW (in percentage) value ranges from a minimum of 12.3% to a maximum of 21.3%, with an average value of 14.59%.

High RDW% results can be used as a proxy for several medical disorders other than anemia, such as persistent inflammation and certain vitamin deficiencies. This test allows participating healthcare facilities to identify at-risk patients sooner and possibly prevent medical emergencies.

The average size or volume of red blood cells in a specific blood sample is determined by measuring the mean corpuscular volume (MCV).

Figure 8 shows the scatter plot between the age of the participants and their MCV value. The mean MCV value was determined to be 81.45, indicating that the MCV measurements for all female participants fell within the defined reference range, with a recorded minimum of 2 and a maximum of 106.

Red blood cell count measures a given amount of red blood cells. Reduced red blood cell count has been associated with several health issues, such as iron or vitamin deficiency anemia, long-term medical conditions, and blood loss. By contrast, conditions including lung disease, dehydration, and polycythemia vera are associated with an increased red blood cell count or erythrocytosis.

The standard reference range for RBC count is established as follows:

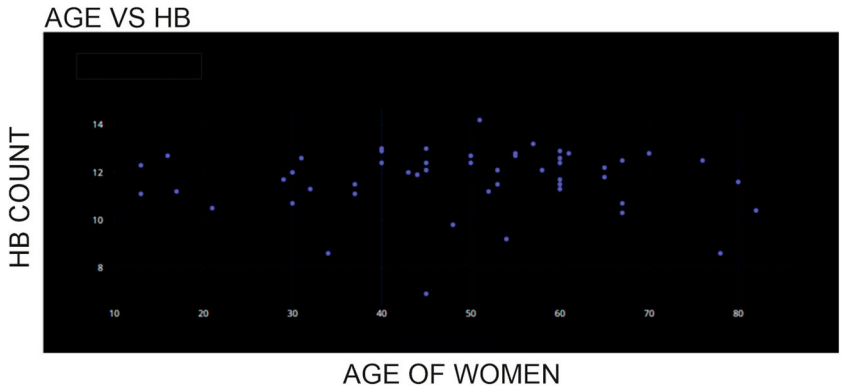

Figure 6. Age V_s Hb count.

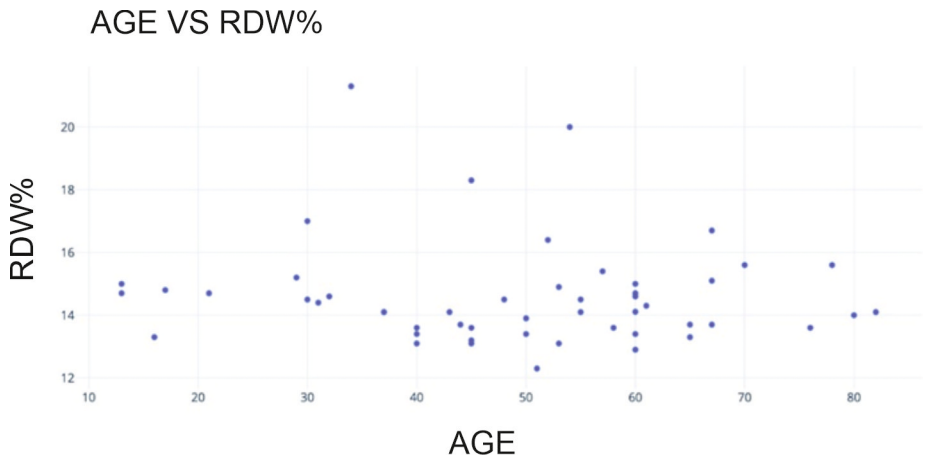

Figure 7. Age Vs RDW% value.

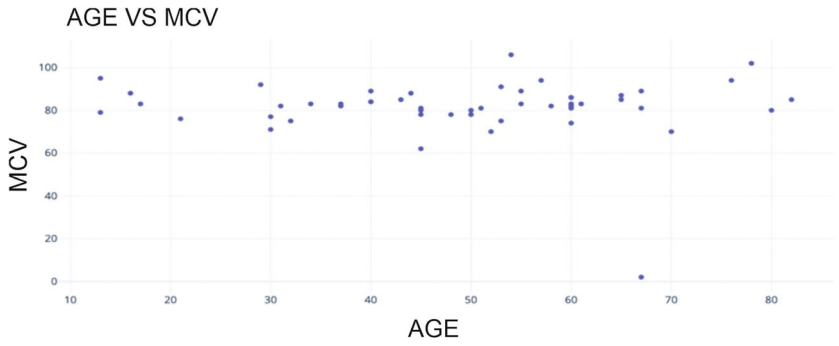

Figure 8. Age Vs MCV value.

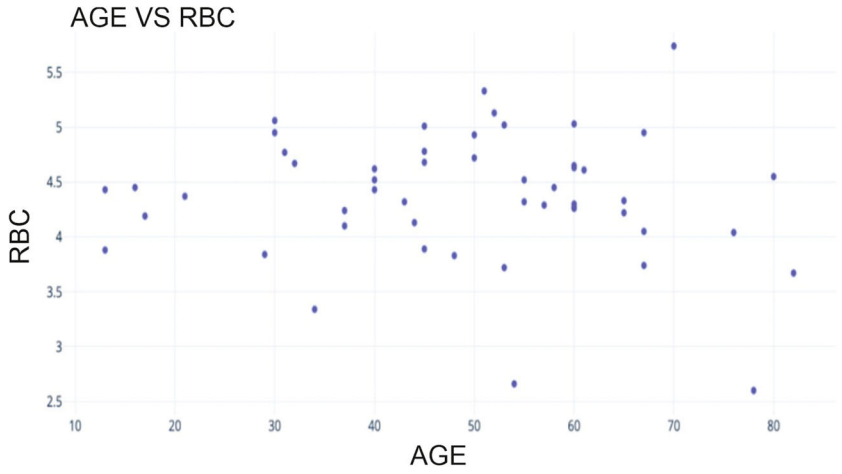

Figure 9. Age Vs RBC value.

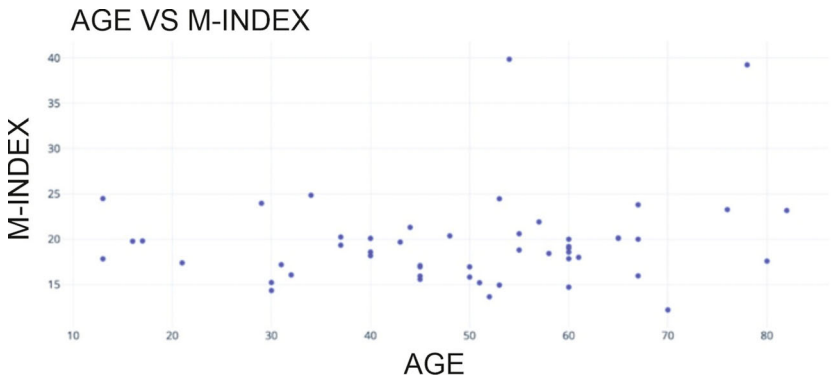

Figure 10. Age Vs M-Index value.

In adult males: A range of 4.5 to 6.0 million red blood cells per microliter.

In adult females, A range of 4.0 to 5.5 million red blood cells per microliter.

Figure 9 shows the scatter plot between the age of the participants and their RBC value. The investigation's findings have shown that the average red blood cell count is 4.37, with a minimum of 2.6 and a maximum of 5.74, with a variance of 0.5.

The Mentzer Index (M-Index) divides the mean corpuscular volume (MCV) by the red blood cell count (RBC). It is used to differentiate between different types of Anemia, especially iron deficiency anemia.

Figure 10 shows the scatter plot between the age of the participants and their M-Index value. Our findings demonstrate that the mean value of the M-INDEX is 19.56, indicating that women in the area under study would have more red blood cells relative to their total red blood cell count. A distinctive observation like this suggests that iron deficiency Anemia, characterized by an inadequate iron supply to facilitate hemoglobin formation, may be observed.

The experimentation gives superior analysis and assessment of different blood parameters. The proposed data analysis and deep learning algorithm achieved average values of hemoglobin count 11.67 g/dL with a 1.33 standard deviation, RDW 14.59%, MCV 81.45, RBC 4.37 per microliter with a variance of 0.5, and M-Index 19.56. Also, the experimental results achieved 97.60% accuracy on 448 × 448 resolution images and 93.41% on 224 × 224 resolution images, demonstrating the effectiveness of the proposed technique for classifying anemia and its subtypes. The experimental results indicate better overlap between

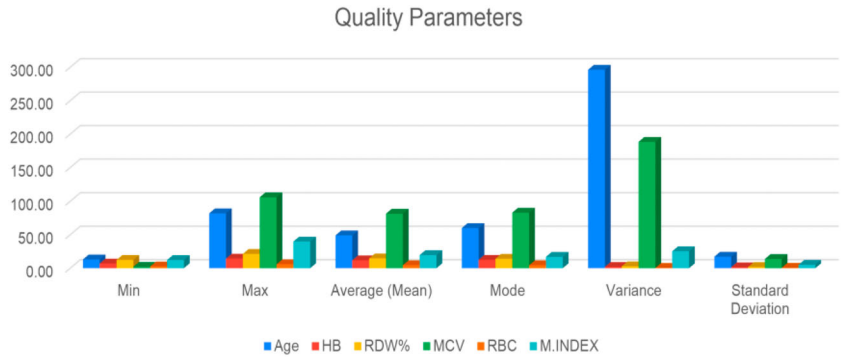

Figure 11. Quality parameters for raw data.

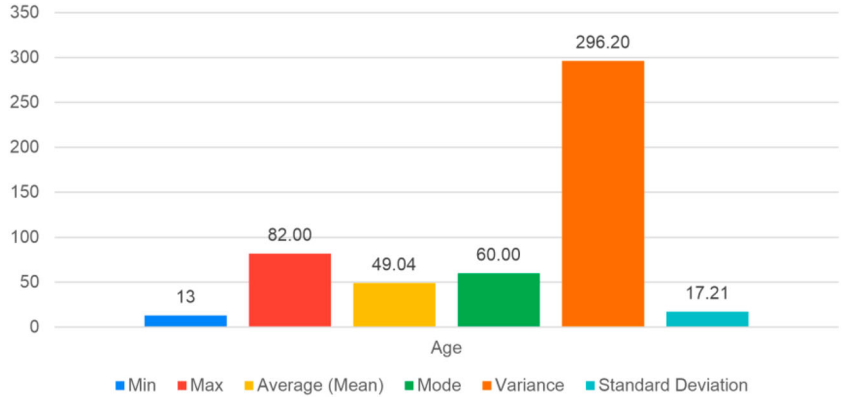

Figure 12. Quality parameters for age.

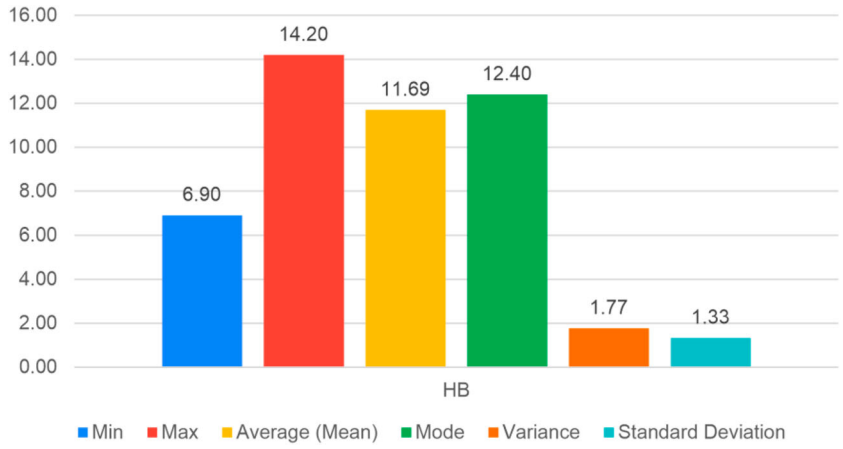

Figure 13. Quality parameters for hemoglobin count.

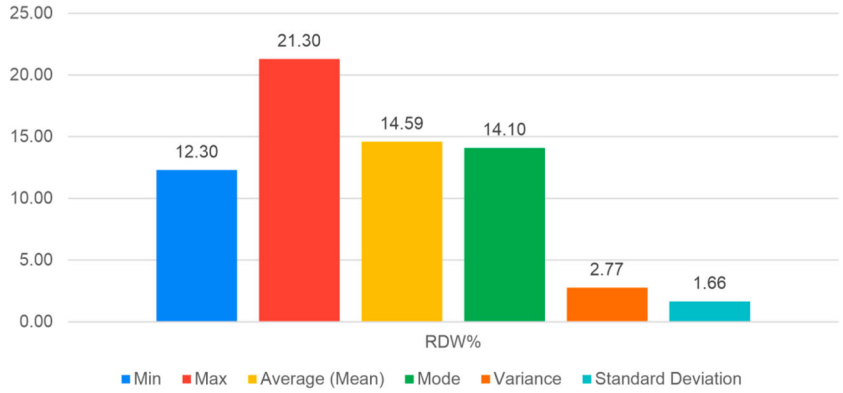

Figure 14. Quality parameters for RDW.

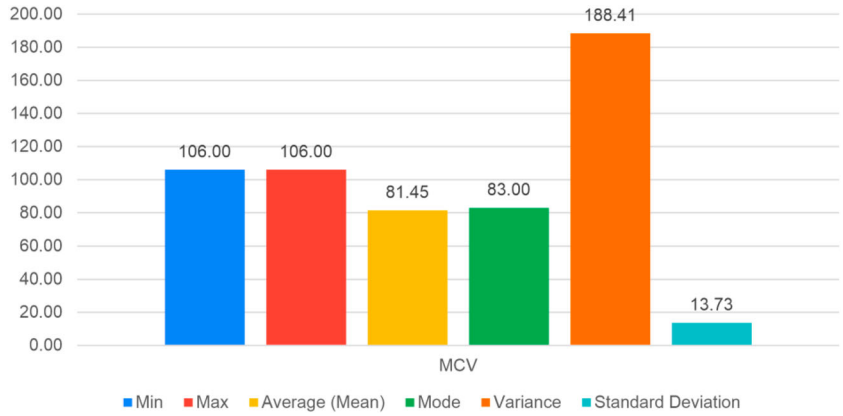

Figure 15. Quality parameters for MCV.

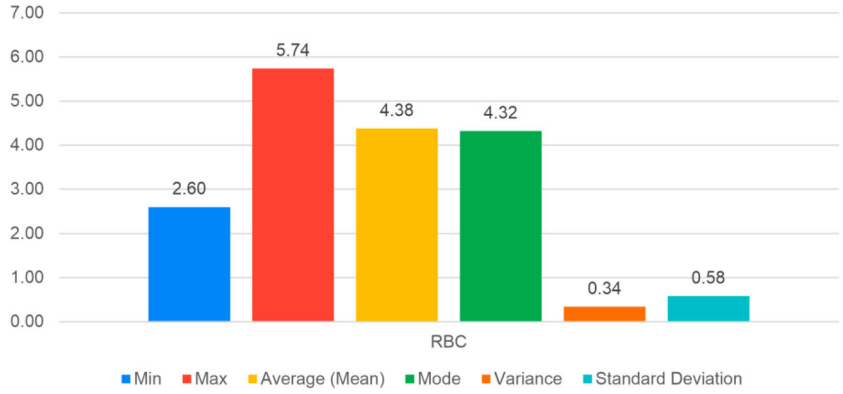

Figure 16. Quality parameters for RBC.

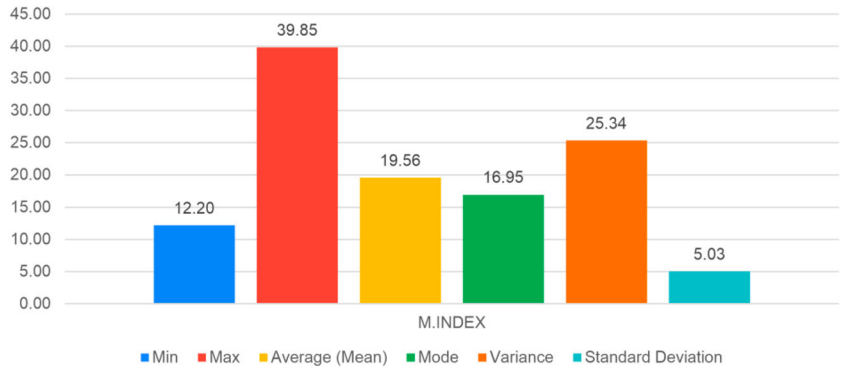

Figure 17. Quality parameters for M-Index.

the automated identification of anemia and manual detection by the experts.

4.1. Comparative analysis

For the proper validation of the proposed system, all the quality parameters such as min (minimum value), max (maximum value), average (mean), mode, variance, and standard deviation for the proposed entity, such as age, hemoglobin (Hb) count, RDW%, MCV, RBC, and M-Index is calculated. For its better analysis, understanding, and interpretation, its graphical representation is shown in Figures 11–17.

5. Conclusion and future work

This study has illuminated various hematological indicators essential for knowing the health state of the women living in the region that was investigated. Analysis of hemoglobin levels, counts of red blood cells, mean corpuscular volume, and the Mentzer Index have yielded important information about the incidence of various forms of anemia, the possibility of an iron shortage, and the morphological features of red blood cells. According to our research, women in the area may be prone to iron deficiency anemia, having more red blood cells than total red blood cells. Furthermore, the examination of MCV values and RBC counts indicates that although the women's RBC counts are typically within the reference range, their red blood cells might be more prominent.

These findings are significant because they may help policymakers and medical professionals understand whether community-based iron deficiency and anemia may be addressed with focused interventions. By identifying these health trends, relevant authorities and healthcare professionals can put necessary measures in place to improve the population's general well-being.

Moreover, using measures such as RDW, RBC count, and MCV has provided an extensive understanding of the fundamental elements influencing fluctuations in hematological parameters. These insights greatly help public health experts and medical professionals trying to create customized healthcare plans to deal with specific health issues in the community.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Ethics approval

We confirmed that the experimental analysis was performed using the free dataset; therefore, ethical approval is Not Applicable.

Data availability statement

The datasets used and/or analyzed during the current study are available from the corresponding author upon request.

ORCID

Nilesh Bhaskarrao Bahadure http://orcid.org/0000-0003-4361-3870
Ramdas Khomane http://orcid.org/0000-0002-8806-7351

References

[1] Asare JW, Appiahene P, Donkoh ET. Detection of anaemia using medical images: a comparative study of machine learning algorithms a systematic literature review. Inf Med Unlocked. 2023;40:1–10. doi:10.1016/j.imu.2023.101283
[2] Appiahene P, Asare JW, Donkoh ET, et al. Detection of iron deficiency anemia by medical images: a comparative study of machine learning algorithms. BioData Min. 2023;16(2):1–20.
[3] Bahadure NB, Ray AK, Thethi HP. Comparative approach of mri-based brain tumor segmentation and classification using genetic algorithm. J Digit Imaging. 2018;31:477–489. doi:10.1007/s10278-018-0050-6
[4] Bahadure NB, Ray AK, Thethi HP. A comparative approach of brain tumor detection using svm, dct and huffman coding in compressed domain. Curr Med Imaging Rev. 3 2018;14:778–787. doi:10.2174/1573405613666170629154727
[5] Hemasri A, Sreenidhi MD, Chaitanya VVK, et al. Detection of rbcs, wbcs, platelets count in blood sample by using deep learning. in 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS). 2023: 47–51.

[6] Gangula Y, KK AM. Detection, classification and counting rbcs and wbcs using deep learning. in 2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC), pp. 512-517, 2023.
[7] Mumford SL, Towler BP, Pashler AL, et al. Circulating microrna biomarkers in melanoma: tools and challenges in personalised medicine. Biomolecules. 2018;8(2). doi:10.3390/biom8020021
[8] Gozdzikiewicz N, Zwolinska D, Polak-Jonkisz D. The use of artificial intelligence algorithms in the diagnosis of urinary tract infectionsmdash;a literature review. J Clin Med. 2022;11(10). doi:10.3390/jcm11102734
[9] Chakraborty S, Kansara K, Kumar RD, et al. Non-invasive estimation of clinical severity of anemia using hierarchical ensemble classifiers. J Med Biol Eng. 2022;42:828–838. doi:10.1007/s40846-022-00750-3
[10] Al-Salmani K, Abbas HH, Schulpen S, et al. Simplified method for the collection, storage, and comet assay analysis of dna damage in whole blood. Free Radical Biol Med. 2011;51(3):719–725. doi:10.1016/j.freeradbiomed.2011.05.020
[11] Schmeiser HH, Muehlbauer K-R, Mier W, et al. Dna damage in human whole blood caused by radiopharmaceuticals evaluated by the comet assay. Mutagenesis. 2019;34(3):239–244. doi:10.1093/mutage/gez007
[12] Kavsaoglu AR, Polat K, Hariharan M. Non-invasive prediction of hemoglobin level using machine learning techniques with the ppg signal's characteristics features. Appl Soft Comput. 2015;37:983–991. doi:10.1016/j.asoc.2015.04.008
[13] Chandra A, Chauhan A, Bansal N, et al. Application of machine learning in hematological diagnosis. in 2021 International Conference on Technological Advancements and Innovations (ICTAI). 2021:665–671.
[14] Nithya R, Nirmala K. Detection of anaemia using image processing techniques from microscopy blood smear images. J Phys Conf Ser. 2022;2318:012043. doi:10.1088/1742-6596/2318/1/012043
[15] Waisberg E, Ong J, Zaman N, et al. A non-invasive approach to monitor anemia during long-duration spaceflight with retinal fundus images and deep learning. Life Sci Space Res (Amst). 2022;33:69–71. doi:10.1016/j.lssr.2022.04.004
[16] Alomar K, Aysel HI, Cai X. Data augmentation in classification and segmentation: a survey and new strategies. J Imaging. 2023;9(2):1–26. doi:10.3390/jimaging9020046
[17] Ahsan MM, Siddique Z. Machine learning-based heart disease diagnosis: a systematic literature review. Artif Intell Med. 2022;128:102289. doi:10.1016/j.artmed.2022.102289
[18] Rikan SB, Azar AS, Ghafari A, et al. Covid-19 diagnosis from routine blood tests using artificial intelligence techniques. Biomed Signal Process Control. 2022;72:103263. doi:10.1016/j.bspc.2021.103263
[19] Zhao Y, Zhang R, Zhong Y, et al. Statistical analysis and machine learning prediction of disease outcomes for covid-19 and pneumonia patients. Front Cell Infect Microbiol. 2022;12:838749. doi:10.3389/fcimb.2022.838749
[20] Asare JW, Appiahene P, Donkoh ET, et al. Iron deficiency anemia detection using machine learning models: a comparative study of fingernails, palm and conjunctiva of the eye images. Eng Rep. 2023;40:1–21.
[21] Appiahene P, Arthur EJ, Korankye S, et al. Detection of anemia using conjunctiva images: a smartphone application approach. Med Novel Technol Devices. 2023;18:100237. doi:10.1016/j.medntd.2023.100237
[22] Yang X, Piety NZ, Vignes SM, et al. Simple paper-based test for measuring blood hemoglobin concentration in resource-limited settings. Clin Chem. 2013;59(10):1506–1513. doi:10.1373/clinchem.2013.204701
[23] Johann KS, Bauer H, Wiegand P, et al. Detecting DNA damage in stored blood samples. Forensic Sci Med Pathol. 2023;19(1):50–59. doi:10.1007/s12024-022-00549-3
[24] Vijayarani S, Sudha S. An efficient clustering algorithm for predicting diseases from hemogram blood test samples. Indian J Sci Technol. 2015;8(17):1–8. doi:10.17485/ijst/2015/v8i17/52123
[25] Boersma E, Vroegindewey MM, van den Berg VJ, et al. Details on high frequency blood collection, data analysis, available material and patient characteristics in biomarcs. Data Brief. 2019;27:104750. doi:10.1016/j.dib.2019.104750
[26] Ahdan S, Setiawansyah S. Android-based geolocation technology on a blood donation system (BDS) using the Dijkstra Algorithm. Int J Appl Inf Technol. 2021;5(1):1–15.
[27] Le HT, Nguyen TTL, Nguyen TA, et al. Bloodchain: a blood donation network managed by blockchain technologies. Network. 2022;2(1):21–35. doi:10.3390/network2010002
[28] Alhazmi L. Detection of wbc, rbc, and platelets in blood samples using deep learning. BioMed Res Int. 2022;2022(Article ID 1499546):1–10. doi:10.1155/2022/1499546
[29] Dimauro G, Griseta ME, Camporeale MG, et al. An intelligent non-invasive system for automated diagnosis of anemia exploiting a novel dataset. Artif Intell Med. 2023;136:102477. doi:10.1016/j.artmed.2022.102477
[30] Haggenmuller V, Bogler L, Weber A-C, et al. Smartphone-based point-of-care anemia screening in rural Bihar in India. Commun Med. 2023;3(38):1–10.
[31] Dimauro G, Camporeale MG, Dipalma A, et al. Anaemia detection based on sclera and blood vessel colour estimation. Biomed Signal Process Control. 2023;81:104489. doi:10.1016/j.bspc.2022.104489
[32] Dhalla S, Maqbool J, Mann TS, et al. Semantic segmentation of palpebral conjunctiva using predefined deep neural architectures for anemia detection. Procedia Comput Sci. 2023;218:328–337. doi:10.1016/j.procs.2023.01.015
[33] Saputra DCE, Sunat K, Ratnaningsih T. A new artificial intelligence approach using extreme learning machine as the potentially effective model to predict and analyze the diagnosis of anemia. Healthcare. 2023;11(5):1–25.
[34] Kistenev YV, Vrazhnov DA, Shnaider EE, et al. Predictive models for covid-19 detection using routine blood tests and machine learning. Heliyon. 2022;8(10):e11185. doi:10.1016/j.heliyon.2022.e11185
[35] Kukar M, Guncar G, Vovko T, et al. Covid-19 diagnosis by routine blood tests using machine learning. Sci Rep. 2021;11:10738. doi:10.1038/s41598-021-90265-9
[36] Chen H, Wang F, Su L, et al. Mathematical statistics of factors affecting the unqualified quality of blood samples in medical examination. in 2020 International Conference on Public Health and Data Science (ICPHDS). 2020: 253–256.
[37] Pfeil J, Nechyporenko A, Frohme M, et al. Examination of blood samples using deep learning and mobile microscopy. BMC Bioinformatics. 2022;23(65):1–14.
[38] Alsheref FK, Gomaa WH. Blood diseases detection using classical machine learning algorithms. Int J Adv

Comput Sci Appl. 2019;10(7):77–81. doi:10.14569/IJACSA.2019.0100712

[39] Noor NB, Anwar MS, Dey M. An effcient technique of hemoglobin level screening using machine learning algorithms. in 2019 4th International Conference on Electrical Information and Communication Technology (EICT). 2019: 1–6.

[40] Golap MA-u, Hashem MMA. Non-invasive hemoglobin concentration measurement using mggp-based model. in 2019 5th International Conference on Advances in Electrical Engineering (ICAEE), pp. 1–6, 2019.

[41] Gun_car G, Kukar M, Notar M, et al. An application of machine learning to haematological diagnosis. Sci Rep. 2018;8(1):1–12.

[42] Narmatha C, Eljack SM, Tuka AARM, et al. A hybrid fuzzy brainstorm optimization algorithm for the classification of brain tumor MRI images. J Ambient Intell Humaniz Comput. 2020;96(01):867–879.

[43] Jiang P, Ergu D, Liu F, et al. A review of yolo algorithm developments. In The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021), pp. 1066–1073, Procedia Computer Science, 2022.