Optimizing Anemia Detection Using Effective Computational Techniques

Swapnil M Parikh[1], Dukhbhanjan Singh[2], Hemal Thakker[3], Murugan R[4]

[1]Computer Science and Engineering, Parul Institute of Technology, Parul University, Vadodara, Gujarat, India
[2]Centre for Research Impact and Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, 140401, Punjab, India.
[3]Department of ISME, ATLAS SkillTech University, Mumbai, Maharashtra, India
[4]Department of Computer Science and Information Technology, Jain (Deemed to be University), Bangalore, Karnataka

Email: swapnil.parikh17761@paruluniversity.ac.in[1], dukhbhanjan.singh.orp@chitkara.edu.in[2], hemal.thakker@atlasuniversity.edu.in[3], murugan@jainuniversity.ac.in[4]

Abstract

Worldwide, anemia is the most common blood disease. The World Health Organization (WHO) defines anemia as the lack of red blood cells, which prevents the body from carrying enough oxygen to satisfy its requirements. Anemia is characterized by decreased erythrocyte mass, blood hemoglobin, and hemocrit levels. Early detection and accurate diagnosis are essential for effective management and therapy. The study's goal is to develop an algorithm for optimizing anemia detection utilizing an effective computational technique. The study proposed a brand-new Dynamic Gannet-tuned Light Gradient Boosting Machine (DG-LGBM) model for the detection of anemia in typical clinical practice settings. In this study, anemia data is collected from a publicly available dataset from Kaggle. The data was preprocessed using data cleaning and normalization for the obtained data. The study aims to improve the predicted accuracy and efficiency of anemia diagnosis by utilizing clinical and biochemical markers. The results demonstrate that, in comparison to traditional methods, the DG-LGBM model performed better in terms of anemia detection rates, highlighting the potential of computational tools to completely transform anemia screening practices. In a comparative analysis, the proposed model is validated using precision (92%), recall (91.71%) f1-score (93.07%), and accuracy (91.06%) values. In addition to advancing the area of medical diagnostics, this study highlights the significance of technology in enhancing healthcare outcomes for impacted communities.

Keywords

Introduction

Anemia is a major worldwide health concern that impacts millions of people from various demographic backgrounds, resulting in severe morbidity and a decreased quality of life (Wekalao et al, 2024). Anemia, which is defined as a lack of red blood cells or hemoglobin, can have several causes, such as dietary inadequacies, long-term illnesses, and hereditary conditions. According to estimates from the World Health Organization (WHO), anemia affects over 1.62 billion people worldwide, with young children, pregnant women, and persons with chronic diseases having the highest frequency (KT et al, 2022). The illness affects one's ability to operate physically but also one's cognitive abilities and general health.

Hematological assays, such as serum ferritin levels and complete blood counts (CBC), are the mainstay of traditional approaches for detecting anemia (Biswas and Rakshit, 2020). Even through these tests are common in clinical settings, their need for specialized tools and individuals with training might make them less accessible, especially in settings with low resources (Garcia-Casal et al, 2023). Furthermore, the interpretation of these data could not always match the clinical images, which could result in an incorrect diagnosis or a postponed course of therapy (Bahadoran et al, 2022). As a result, novel, effective, and widely available techniques for the detection of anemia are desperately needed to enable prompt diagnosis and treatment.

This study aims to develop a novel Dynamic Gannet-tuned Light Gradient Boosting Machine (DG-LGBM) model and an efficient algorithm for the identification of anemia cases. The following sections provide a list of literary reviews, the methodology, the results, and the conclusion.

Methodology

The research used the Kaggle dataset for anemic detection and preprocessing using data cleaning and min-max normalization to obtain the data. In the study, a Dynamic Gannet-tuned Light Gradient Boosting Machine (DG-LGBM) model was created to enhance anemia detection.

1. Dataset

We gathered the Kaggle source dataset. (https://www.kaggle.com/datasets/zeesolver/uhygtttt/data). The dataset contains 104 images of the conjunctiva of the eyes, together with information on the hemoglobin level and anemia. Subsequently, preprocessing was performed on the images, including cropping and extracting the proportion of red, green, and blue pixels. Number: A unique number assigned to every dataset element. Sex: The person's gender (e.g., Male or Female). %Red Pixel: The proportion of red pixels in the linked picture. %Green Pixel: The proportion of green pixels inside the linked picture. % Blue Pixel: The proportion of blue pixels in the linked picture. Hb: The amount of hemoglobin expressed in grams per deciliter (g/dL). Anemia: A measure of anemia status (0 being non-anemia, 1 being anemic).

2. Preprocessing using Data cleaning and Normalization

Cleaning and normalizing data is a part of data preprocessing, which improves the quality of data. Normalization brings data values into compliance with a common scale, while data cleaning eliminates errors and inconsistencies. All these procedures work together to increase the efficiency of models and analyses, guaranteeing trustworthy information and promoting improved decision-making across a range of applications.

2.1. Data cleaning

The first basic data cleaning activity is to identify inadequate or infeasible values for each column and then proceed to remove or amend it. Additionally, there should be no repetition of entries in the "Number" box. It was recommended that, for all future cases, the entries under "Sex" should reflect only "Male" or "Female". As regards the checklists, that the values in the "%Red Pixel," "%Green Pixel," and "%Blue Pixel" fields do not exceed the limits of 0 % to 100 % and any numeration that seems mistaken. Ensure that "Hb" has represent normal hemoglobin level values; discard any unrealistic values. The final dataset should accurately represent "Anemia" values ensuring they are binary (i.e., 0 or 1), and correct any values that do not meet this criterion.

2.2. Min-Max Normalization

It was crucial to normalize the data if we do not want complexes of errors in the input data lead to obtaining improper results. A mathematical function of converting numerical data through the elimination of variation from the data set and then reconstructing the variability into another form of the data set. The use of a min-max normalization framework makes it possible. In today's normalization mainly three methods are used for normalization, those are range normalization, min-max normalization, and standardization method, among all these min-max normalization was the most commonly used method. The values in the given dataset are scaled within the given minimum and maximum values and then every value was transformed by the use of equation (1).

$$u' = \frac{u - min_B}{max_B - min_B} \left(\frac{new_{max_B}}{-new_{min_B}} \right) + new_{min_B} \qquad (1)$$

Where

B: Attribute data

max_B, min_B: A's absolute, minimum, and maximum value.

u': The updated value of each data item.

u: The previous value of each data item.

new_{max_B} , new_{min_B}: Max and min value of the range.

3. Dynamic Gannet tuned Light Gradient Boosting Machine (DG-LGBM)

An enhanced hybrid model for the diagnosis of anemia is the DG-LGBM. This method improves prediction accuracy and flexibility by fusing the sturdy LGBM structure with the ideas of the Gannet optimization algorithm. The Gannet method dynamically optimizes hyper parameters, allowing the DG-LGBM to navigate complicated datasets with efficiency that are frequently encountered in medical diagnostics. Detecting subtle patterns linked to anemia is made easier with this hybrid model, which makes use of LGBM's capacity to handle large-scale data and enhance learning efficiency. The DG-LGBM is capable of precisely identifying instances of anemia and enabling prompt therapies by evaluating a range of clinical variables, including blood counts and biochemical indicators. This novel technique has a great deal of potential to enhance

anemia detection procedures, providing a trustworthy instrument for medical practitioners in their diagnostic procedures.

3.1. Light Gradient Boosting Machine (LGBM)

The LGBM is an open-source promotional goal for the gradient-boosting decision tree method. LGBM was one of the fastest and most efficient algorithms. The method was used for sorting, regression, classification, and several other machine-learning applications, and it allows for efficient parallel training. Using a histogram, the LGBM technique uses a conservative growth strategy with depth restrictions, minimizes memory use, and speeds up training compared to Xtream Gradient Boosted (XGBoost).

The histogram will have a width of k bins if the linearity of the floating-point eigenvalue is limited to k bins. The fundamental notion behind LGBM utilizing a histogram, LGBM can hold 8-bit integers, reduces memory use to 1/8 of the original, and does not require a large amount of storage for pre-sorted results. The precision of the LGBM mode was not affected by this imprecise partition. There are three phases in the LGBM boosting type. X was available as an easily navigable, pre-processed gathering data collection.

Step 1. Initialize by (2), the weak learner.

$$e_0(w) = arg\min_d \sum_{j=1}^n K(z_j, d) \tag{2}$$

Where $e_0(w)$ the weak learner basis was function and $K(z_j, d) = K(z, e(w)) = (z - e(w))^2$ with n being the number of samples and being the loss function.

Step 2: Iteratively calculate M times for weak learners

a. Determine the loss function's negative gradient as it was evaluated in the present model in (3) for the sample $w_j \in W \forall j = 1,2,....,m$

$$q_{mi} = -\left[\frac{\partial K(z_j, e(w_j))}{\partial e(w_j)}\right] e(w) = e_{n-1}(w) \tag{3}$$

Where q_{mi} the loss function of the negative gradient

b. The residual q_{mi} value obtained was considered the sample's updated actual value. Construct a regression tree to:
$\{(w_1, q_{n1}),...., (w_m, q_{nm})\}$ can generate a new regression tree $e_n(w)$

c. Using d_{mj} in (4) as a linear procedure to forecast a leaf by determining the ideal value of leaf area node sector value for lowering the loss function $i = 1,2,....,I$.

$$d_{mj} = \underset{d}{\text{argmin}} \sum_{w_j = Q_{mj}} K\left(z_j, e_{n-1}(w_j + d)\right), j = 1,....,N \tag{4}$$

d. equation (5), update the robust learner.
$$e_n(w) = e_{n-1}(w) + \sum_{i=1}^I d_n J(w \in Q_{mj}) \tag{5}$$

Where J was the indicator function, $e_{n-1}(w)$ was the pre-weak learner, and $e_n(w)$ was the current weak learner.

Step 3: Using (6), find the final regression tree.

$$E(w) = \sum_{n=1}^{N} \sum_{i=1}^{I} d_n J(w \in Q_{mj})$$ (6)

The normalized total decrease of criteria that a feature brings about its relevance was determined. Another name was the $Gini$ significance. $Gini\ (p)$ is the representative for $Gini$.

$$Gini(o) = \sum_{k=1}^{K} O_k(1 - o_k) = 1 - \sum_{l=1}^{K} o_k^2$$ (7)

Where o_l is the weight of the l-label and K is the number of labels.

3.2. Dynamic Gannet (DG)

The DG was a metaheuristic optimization method that drew inspiration from evolution as a natural process. It was mostly driven by the gannets' (a particular kind of bird) innate predation traits. It uses a population of solutions to repeatedly develop toward optimal or nearly ideal solutions, emulating the concepts of genetics and survival of the fittest. There are two steps to the optimization algorithm: exploration and exploitation. Four stages make up the mechanism: abrupt rotation, random wandering, U-shaped dive mode, and V-shaped dive mode.

Natural habits of gannet: The gannet appears to have a thin neck, be fat, and be stubby. It lives near lakes or on the seaside and feeds on fish. Because of its keen sight, it can locate prey even when it is in flight. The steps that follow will enable to find the ideal response.

Step 1: Initialization: A population of possible solutions, called individuals, was created at random. Every person stands for a potential fix for the optimization issue. Equation (8) illustrates how the starting population, which was believed to be the gannets, was randomly started with boundary values.

$$m_{o,r} = c_1 \times (upA_r - l\omega A_r) + l\omega A$$ (8)

Here, $m_{o,r}$ designates the location of the oth search agent in the rth dimension. Next, upA and IwA, respectively, are used to indicate the upper and lower boundaries. Ultimately, as c_1 illustrates, the random number falls between 0 and 1. The memory matrix in the operation was specified by the word YT. The memory matrix was supposed to represent the changes in gannet positions during the iteration. The changes in the matrix value are based on the fitness assessment.

Stage 2: Exploration: In the phase, gannets scour the water's surface in search of prey. Depending on the water's depth, they can modify their dive patterns after locating the prey. As stated in Equations (9) and (10), two different kinds of dives have been found in nature: U-shaped dives and V-shaped dives.

$$v = 2 * \cos(2 * \pi * qc_2) * m$$ (9)

$$u = 2 * W(2 * \pi * qc_3) * m$$ (10)

$$m = 1 - \frac{md}{nm} \tag{11}$$

The present iteration and highest iteration count are indicated by the variables md and mn in Equation (11). Furthermore, as stated in qc_2 and qc_3, respectively, the random values fall between 0 and 1. Equation (12) was used to update and determine the new location with the aid of these two diving criteria.

$$H_i(m + 1) = \begin{cases} H_i(m) + b1 + b2, & o \geq 0.5 \\ H_i(m) + a1 + a2, & o < 0.5 \end{cases} \tag{12}$$

$$b2 = V * \left(H_i(m) - H_c(m)\right) and\ a2 = U * \left(H_i(m) - H_b(m)\right) \tag{13}$$

Equation (14) represents the terms V and U from Equation (12).

$$V = (2 * qc_4 - 1) * v, \quad U = (2 * qc_5 - 1) * u \tag{14}$$

The variables qc_4 and qc_5 in the equation above represent random values between 0 and 1. A random sample of the population was denoted by $H_c(m)$, the average location was indicated by $H_b(m)$, and the current search agent was designated by $H_i(m)$. Here, Equation (15) was used to calculate the average position.

$$H_b(m) = \frac{1}{N}\sum_{i=1}^{N} H_i(m) \tag{15}$$

Stage 3: Exploitation: It takes much more energy to capture the prey once the searchers dive into the water. As a result, when the fish struggles to break free of the surface, the gannets may display incredible force. When the gannet has enough energy, it can capture enough prey. Equation (16) yields the new location based on the chosen capturing tactic.

$$H_i(m + 1) = \begin{cases} m * \delta * \left(H_i(m) - H_{bt}(m)\right) + H_i(m). & Cp \geq s \\ H_{bt}(m) - \left(H_i(m) - H_{bt}(m)\right) * R * m, & Cp < s \end{cases} \tag{16}$$

The constant parameter in the case was denoted by s, which has a set value of 0.2. The following equations are used to determine the term Cp in the equation above.

$$Cp = \frac{1}{O * m2} \tag{17}$$

$$m2 = 1 + \frac{md}{mn} \tag{18}$$

$$O = \frac{nb * uz^2}{k} \tag{19}$$

$$k = 0.2 + (2 - 0.2) * qc_6 \tag{20}$$

The range from 0 to 1, as shown in qc_6, was contained in the random value here. Next, it is stated that the gannet's mass and velocity, designated as nb and uz, are 1.5 m/s and 2.5 kg, respectively. Equation (21) was also used to evaluate the term δ.

$$\delta = Cp * |H_i(m) - H_{bt}(m)| \tag{21}$$

$$R = Luy(C) \tag{22}$$

The DG technique was modified and built upon to address the particular challenges related to pooling center locations and transport optimization in the engine components industry. The changes involve adding more parameters, modifying the selection procedure, and improving the genetic processes to better fit the features of the case. It hopes that a thorough description of the DG technique will improve readers' comprehension of the fundamental optimization strategy that underpins the suggested Modified Dynamic Gannet (MDG). To determine the ideal value that improves performance, the recently created MDG was used.

The traditional DG approach creates a new equation that inherits. The DG demonstrates some essential points, including a faster rate of convergence, the capacity to tackle the local optima problem, and strong processing power. However, it has to be improved because giving the solution a larger number of parameters tends to make it more difficult. Because of the issues with parameter setup and adaptation, which are mostly related to determining the delta parameter as given by Equation (16). Equation (23) illustrates how the new formulation, which takes the place of Equation (16) addresses such problems. A new algorithm, MDG, has been proposed to achieve this. Equation (26) illustrates the new mathematical statement in MDG, which follows Equation (23).

$$\delta = {}^{\beta}/_{\gamma} \tag{23}$$

$$\gamma = {}^{\beta}/_{\alpha} \tag{24}$$

The estimates for the terms α and β in the equation are as follows:

$$\alpha = {}^{N}/_{100} \tag{25}$$

$$\beta = {}^{N}/_{20} \tag{26}$$

The total population considered is indicated by the word. Therefore, the greatest outcome is determined by Equation (16), which uses the new delta value.

Results and Discussion

A laptop running Windows 11 with an Intel i5 8th Gen CPU, 32 GB of RAM, and a Python 3.10.1 environment was used to test the recommended solution. The section covers the experimental setup and conditions as well as the efficacy of the suggested approach. Compared to the current methods, which include Vision Transformer (ViT) (Appiahene et al, 2023), Visual Geometry Group (VGG16) (Appiahene et al, 2023), and Residual networks (ResNet50) (Appiahene et al, 2023). Table 1 displays the overall results.

1. Accuracy and Loss

Accuracy in the context of anemia detection was the ratio of the model's accurate predictions to its total number of predictions. A high accuracy means that the model can distinguish between anemic and non-anemic instances with good precision. Loss, on the contrary, evaluates how much the actual values during training departed from the expected values, which aids in the

optimization process. Reducing loss was essential for enhancing model performance, which in turn improves patient anemia diagnosis accuracy. Figure 1 shows (A) Accuracy and (B) Loss.
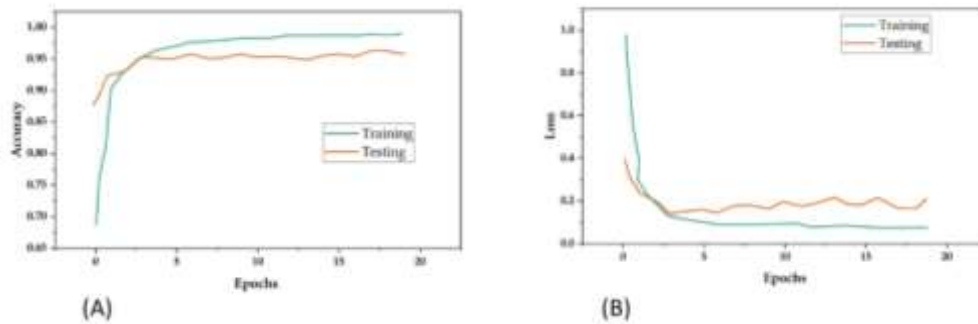

Figure 1.   Result of (A) Accuracy and (B) loss

2. Accuracy

Accurate identification of anemia was essential for prompt diagnosis and therapy. The application of cutting-edge techniques and technology improves sensitivity and specificity, ensuring accurate diagnosis of anemia across a range of clinical contexts and demographics. Figure 2 shows a comparison of the accuracy.
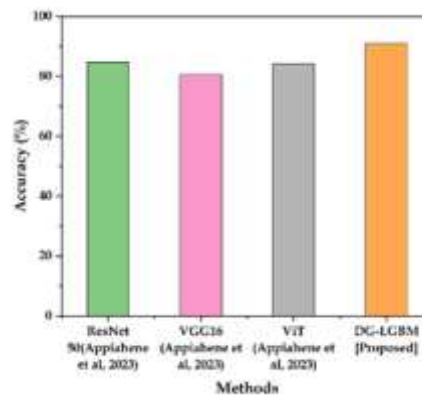

Figure 2.   Result of Accuracy

The current techniques ResNet50, VGG16, and ViT obtained (84.79%), (80.56%), and (84.08%), respectively, and the accuracy reached by the DG-LGBM methodology was 91.06%. The outcome demonstrates that the suggested strategy performs significantly better than the current techniques.

3. Precision

The process of diagnosing anemia entails determining low hemoglobin levels using precision diagnostic procedures, such as blood tests and sophisticated computational approaches. This guarantees an accurate diagnosis and prompt intervention for efficient patient care. A comparable precision query was shown in Figure 3.
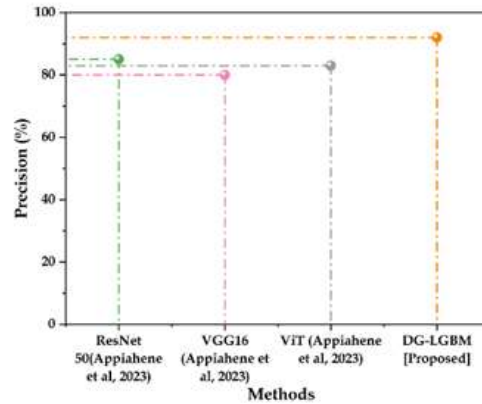
Figure 3.   Result of Precision

The DG-LGBM strategy yielded a precision of 92%, compared to 85%, 80%, and 83% for the existing techniques of ResNet50, VGG16, and ViT. The outcomes demonstrate that the proposed strategy performs noticeably better in terms of precision than the existing approaches.

4. Recall

The anemia detection was to precisely identify anemic patients by the application of efficient computer methods. The high recall value was critical for the identification of many actual positive cases as possible by the applied approach, excluding cases of missed anemia diagnosis, which will enhance the efficiency of anemia screening campaigns. Figure 4 depicts a similar recall problem.
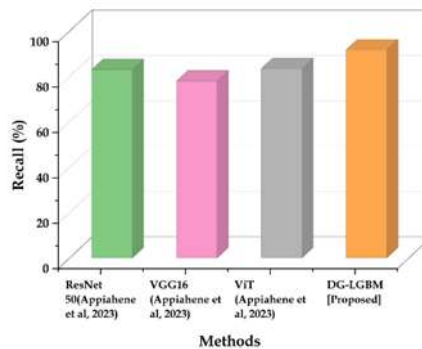

Figure 4.   Result of Precision

A recall of 91.71% was attained by the DG-LGBM methodology compared to 83%, 78%, and 83.5% by the existing techniques of ResNet50, VGG16, and ViT. In terms of recollection, results show that the suggested alternative performs significantly better than the existing methods.

5. F1-Score

Thus, the f1-score among other measures of accuracy and recall was crucial in anemia identification. Optimizing the system of determining the effectiveness of diagnostic models and considering both true positives and false negatives, increases the accuracy of diagnosing anemia in patients. The f1 scores are compared in figure 5.
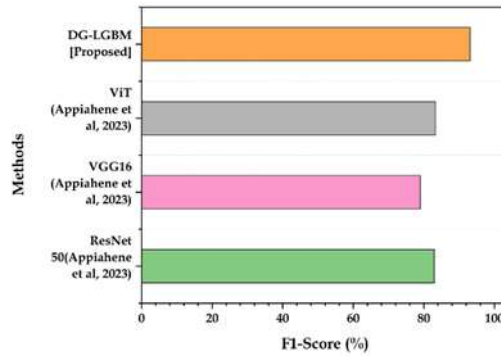
Figure 5.   Result of Precision

The F1-score of the DG-LGBM approach was 93.07%, outperforming the F1-scores of the existing techniques ResNet50, VGG16, and ViT, which were 83%, 79%, and 83.3%, respectively. The results show that the suggested strategy performs much better in terms of F1score than the existing approaches.

Table 1. Overall performance

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ResNet50 (Appiahene et al, 2023)	84.79	85	83	83
VGG16 (Appiahene et al, 2023)	80.56	80	78	79
ViT (Appiahene et al, 2023)	84.08	83	83.5	83.3
DG-LGBM [proposed]	91.06	92	91.71	93.07

References

Aliyu, H. A., Razak, M. A. A., Sudirman, R., & Ramli, N. (2020). A deep learning AlexNet model for classification of red blood cells in sickle cell anemia. *International Journal of Artificial Intelligence, 9*(2), 221-228. https://doi.org/10.11591/ijai.v9.i2.pp221-228

An, R., Huang, Y., Man, Y., Valentine, R. W., Kucukal, E., Goreke, U., Sekyonda, Z., Piccone, C., Owusu-Ansah, A., Ahuja, S., & Little, J. A. (2021). Emerging point-of-care technologies for anemia detection. *Lab on a Chip, 21*(10), 1843-1865. https://doi.org/10.1039/D0LC01235A

Appiahene, P., Chaturvedi, K., Asare, J. W., Donkoh, E. T., & Prasad, M. (2023). CP-AnemiC: A conjunctival pallor dataset and benchmark for anemia detection in children. *Medicine in Novel Technology and Devices, 18*, 100244. https://doi.org/10.1016/j.medntd.2023.100244

Bahadoran, M., Seyfari, A. K., Sanati, P., & Chua, L. S. (2022). Label free identification of the different status of anemia disease using optimized double-slot cascaded microring resonator. *Scientific Reports, 12*(1), 5548. https://doi.org/10.1038/s41598-022-09504-2

Biswas, U., & Rakshit, J. K. (2020). Detection and analysis of hemoglobin concentration in blood with the help of photonic crystal based micro ring resonator structure. *Optical and Quantum Electronics, 52*, 1-16. https://doi.org/10.1007/s11082-020-02566-4

Dutta, M., Bhise, M., Prashad, L., Chaurasia, H., & Debnath, P. (2020). Prevalence and risk factors of anemia among children 6–59 months in India: A multilevel analysis. *Clinical Epidemiology and Global Health, 8*(3), 868-878. https://doi.org/10.1016/j.cegh.2020.02.015

Garcia-Casal, M. N., Dary, O., Jefferds, M. E., & Pasricha, S. R. (2023). Diagnosing anemia: Challenges selecting methods, addressing underlying causes, and implementing actions at the public health level. *Annals of the New York Academy of Sciences, 1524*(1), 7-50. https://doi.org/10.1111/nyas.14996

Jain, P., Bauskar, S., & Gyanchandani, M. (2020). Neural network based non-invasive method to detect anemia from images of eye conjunctiva. *International Journal of Imaging Systems and Technology, 30*(1), 112-125. https://doi.org/10.1002/ima.22359

Khaleduzzaman, K., Mahmud, M. H., & Podder, P. K. (2021). Detection and implementation of blood group and Hb level by image processing techniques. *Australian Journal of Engineering and Innovative Technology, 3*(5), 73-81. https://doi.org/10.34104/ajeit.021.073081

Kilicarslan, S., Celik, M., & Sahin, Ş. (2021). Hybrid models based on genetic algorithm and deep learning algorithms for nutritional anemia disease classification. *Biomedical Signal Processing and Control, 63*, 102231. https://doi.org/10.1016/j.bspc.2020.102231

NK, N., Prasad, K., & Singh, B. M. K. (2022). Analysis of red blood cells from peripheral blood smear images for anemia detection: A methodological review. *Medical & Biological Engineering & Computing, 60*(9), 2445-2462. https://doi.org/10.1007/s11517-022-02614-z

Muthanna, F. M., Karuppannan, M., Abdulrahman, E., Uitrakul, S., Rasool, B. A. H., & Mohammed, A. H. (2022). Prevalence and associated factors of anemia among breast cancer patients undergoing chemotherapy: A prospective study. *Advances in Pharmacological and Pharmaceutical Sciences, 2022*(1), 7611733. https://doi.org/10.1155/2022/7611733

Shahzad, M., Umar, A. I., Shirazi, S. H., Khan, Z., Khan, A., Assam, M., Mohamed, A., & Attia, E. A. (2022). Identification of anemia and its severity level in a peripheral blood smear using 3-tier deep neural network. *Applied Sciences, 12*(10), 5030. https://doi.org/10.3390/app12105030

Viramgami, A., Soundarajan, S., Sheth, A., & Upadhyay, K. (2022). Validity of point-of-care device for diagnosing anemia in workers exposed to lead. *Indian Journal of Community Health, 34*(4), 573-576. https://doi.org/10.47203/IJCH.2022.v34i04.023

Wekalao, J., Mandela, N., Apochi, O., Lefu, C., & Topisia, T. (2024). Nanoengineered graphene metasurface surface plasmon resonance sensor for precise hemoglobin detection with AI-assisted performance prediction. *Plasmonics*, 1-24. https://doi.org/10.1007/s11468-024-02489-w

Yıldız, T. K., Yurtay, N., & Öneç, B. (2021). Classifying anemia types using artificial learning methods. *Engineering Science and Technology, an International Journal, 24*(1), 50-70. https://doi.org/10.1002/ima.22359