Identifying Early Anemia Using Machine Learning Algorithm

Shilpa Priya
M.Tech Student
Dept of CSE, SIET, Tumkur
kavya26feb@gmail.com

Dr. Basavesha D
Associate Professor
Dept of CSE, SIET, Tumkur
basavesha@gmail.com

Dr. Girish L
Associate Professor
Dept of AI&DS, SIET, Tumkur
girishltumkur@gmail.com

Abstract

This study explores the relationship between between Interleukin-6 (IL6) and Interleukin-8 (IL8) cytokines and auto-immune reactions in Sickle Cell Anemia (SCA) patients, aiming to predict their presence based on genetic factors like Haptoglobin alleles using artificial neural networks. The study, conducted on 60 SCA patients and 74 healthy individuals, found a significant association between Haptoglobin alleles and IL6/IL8 production, achieving an accuracy of 90.9% and an r-squared value of 0.88. Concurrently, the broader context of anemia as a global health issue, particularly affecting mothers and children, underscores the importance of non-invasive detection methods,like those based on machine learning and deep learning techniques. These methods, exemplified by convolutional neural networks (CNNs) in blood analysis, offer efficient and cost-effective avenues for early diagnosis and treatment of anemia, highlighting the pivotal role of artificial intelligence in healthcare advancements.

I. Introduction

A reduction in hemoglobin concentration is indicative of anemia below a specified threshold, which varies based on factors such as gender, age, and physiologicall condition, smoking habits, and altitude of the population under evaluation. Current recommendations suggest hemoglobin cut-offs ranging 11.6 to 12.3 g/dl in women and 13 to 14.2 g/dl in males. Severe anemia often stems from trauma, undernourishment, parasite diseases, or underlying medical conditions or medical conditions like gastrointestinal bleeding. In emergency settings like the ER, acute blood loss from conditions such as trauma or gastrointestinal bleeding can lead to severe anemia, necessitating swift identification and restoration of blood volume to prevent hemorrhagic shock and potential fatalities. Classic symptoms of anemia include fatigue, shortness of breath, pale mucous membranes, and resting tachycardia. Prior studies have indicated that hemoglobin is capable of absorbing green light. and reflects red light, influencing tissue coloration. [1]

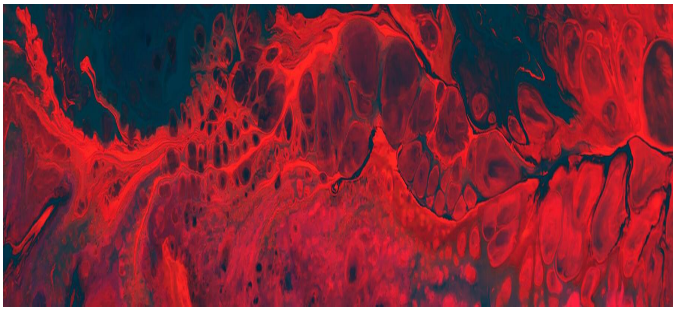


Fig. 1. Sample Anemia Blood Structure

The comparative study on machine learning algorithms for detecting iron deficiency anemia revealed significant advancements in the accuracy and efficiency of anemia detection through the utilization of medical images. By analyzing palpable palm images from Ghana, the authors demonstrated the capabilities of artificial intelligence models in aiding healthcare professionals in diagnosing anemia. The study showcased the importance of data collection and algorithm analysis, including input from several authors focusing on different aspects of the research. The datasets employed inside the inquiry were made publicly available on the Mendeley Data repository, ensuring transparency and reproducibility in the scientific community. Overall, the findings suggest a promising future for the incorporation of machine learning and medical imaging technologies in improving the diagnosis and management of anemia. [2]

Diabetes, renal syndrome, cancer, HIV/AIDS, inflammatory bowel disease, and cardiovascular disease are among the illnesses associated with the complex prevalence of anaemia. Other significant causes include hemoglobinopathies, bilharzia, and malaria [8,9]. Iron deficiency, sickle cell disease, thalassaemia, aplastic anaemia, and vitamin or iron deficiency are only a few of the different types of anaemia that exist. There are numerous causes for every form of anaemia, which range from mild to severe and can be either transient or long-term [2]. The laboratory procedure for diagnosing and detecting anaemia in response to clinical concerns has many challenges in practice, including insufficient funding for medical tests and insufficient technical. [3]

Anemia, a widespread nutritional deficiency disorder, poses a significant worldwide health issue that has an impact on people in both developed and developing nations. It is characterized by a low concentration of Red Blood Cells (RBCs) or Hemoglobin (Hb) in the bloodstream. According to the World Health Organization (WHO), anemia is defined as a condition where the quantity of red blood cells or their ability to transport oxygen is insufficient to meet the body’s physiological demands.[4]

A hemoglobin concentration below a predetermined cut-off value, which varies depending on the population being evaluated’s age, gender, physiological state, smoking status, and altitude, is what is known as anemia. The current recommended ranges for hemoglobin cut-off are 11.6 to 12.3 g/dL for women and 13 to 14.2 g/dL for men. Severe anemia can be brought on by trauma or other illnesses such gastrointestinal bleeding, but it also frequently arises from malnutrition, parasite infections, or underlying ailments. Acute blood loss illnesses, such as trauma or gastrointestinal hemorrhage, can result in severe anemia in emergency rooms. In order to preserve patients, it is critical to identify these conditions quickly and restore circulatory volume as soon as possible. [5]

Anemia is a serious global health issue defined by the World Health Organization (WHO) as having hemoglobin (Hb) levels below 12.0 g/dL in females and 13.0 g/dL in males . Iron deficiency anemia (IDA) is the most common nutritional deficiency worldwide, affecting about 30% of people, with gastrointestinal bleeding, decreased dietary iron, and impaired iron absorption as key factors. The causes of anemia are diverse, including iron and vitamin deficiencies, hemolytic anemia, aplastic anemia, sickle cell anemia, thalassemia, chronic diseases, and nutritional deficiencies like iron, vitamin A, B vitamins, and folic acid, as well as chronic inflammation, parasitic infections, and congenital conditions . Diagnosing anemia traditionally involves drawing a blood sample and analyzing its hemogram, but this method is invasive, painful, and challenging, especially for pediatric groups. [6]

Artificial intelligence and machine learning are the foundational pillars of a new revolution in computing. They enable the identification of patterns and the making of relevant predictions about the future. Deep learning,a portion of machines learning, is a field that focuses on learning and improving autonomously through the examination of computer algorithms. While machine learning employs simpler concepts, deep learning utilizes artificial neural networks designed to mimic human thinking and learning processes. Deep learning has significantly contributed to image classification, language translation, and speech recognition, addressing various pattern recognition challenges without human intervention. Deep Neural Networks (DNNs) consist of layers capable of complex operations like representation and abstraction, enabling the interpretation of images, sound, and text. These networks, akin to the human brain’s neurons, connect nodes across layers, with deeper networks containing more layers. However, deep learning systems necessitate large datasets for accurate results, requiring substantial amounts of data input. In our project, we specifically focus on image data, employing Convolution-Based Neural Systems (CNNs), where convolution layers with convolutional kernels perform pixel-wise operations, pooling layers summarize features, dense layers form deeply connected networks, and an output layer generates the final output.[7]

age	49	64	43	69
cp	1	3	2	0
trestbps	120	150	172	135
chol	239	219	283	233
fbs	0	1	0	1
thalach	178	163	174	114
exang	0	1	0	1
old peak	1.4	0.6	1.8	0.8
thal	1	2	1	2
target	0	1	0	1

TABLE I
OVERVIEW OF THE DATASET

II. LITERATURE SURVEY

The table provides a dataset comprising several attributes related to individuals, likely from a medical or cardiovascular context. Firstly, it includes the ages of individuals, ranging from 43 to 69 years. The ”cp” column categorizes chest pain types, denoted by values from 0 to 3, representing different levels or kinds of discomfort in the chest experienced. Resting blood pressure (”trestbps”) is indicated in millimeters of mercury (mm Hg) and varies from 120 to 172 mm Hg across the dataset. Cholesterol levels (”chol”) are measured in milligrams per deciliter (mg/dL), ranging from 219 to 283 mg/dL. The ”fbs” column signifies blood sugar levels during fasting , with 0 denoting normal levels and 1 indicating elevated levels. ”Thalach” represents the highest heart rate that was attained during exercise, expressed in beats per minute (bpm) and ranging from 114 to 178 bpm. The existence of exercise-induced angina (”exang”) is indicated by 0 for its absence and 1 for its presence. ”Old peak” refers to the ST depression induced by exercise relative to rest, with values ranging from 0.6 to 1.8. The ”thal” column denotes the thalassemia type of individuals, numerically categorized. Finally, the ”target” column signifies the presence (1) or absence (0) of heart disease in individuals, potentially serving as a target variable for predictive modeling or analysis. These attributes collectively provide a comprehensive view of factors related to cardiac health and potential indicators of heart disease within the dataset.[8]

The literature survey outlined in the provided PDF file explores the identification of iron deficiency anemia using medical images. The study conducts a comparative analysis of various CNN and other machine learning algorithms, k-NN, Decision Tree, Naïve Bayes, and SVM, for anemia detection based on palm images. Gathering datasets from multiple hospitals in Ghana, the research underscores the significance of image preprocessing techniques and dataset augmentation methods. Key steps in the analysis comprise segmenting the Region of Interest (ROI) in images and converting them to the CIELAB color space model. Ethical considerations are paramount, with the study obtaining necessary approvals and providing training sessions for medical personnel involved in data collection.In summary, this research demonstrates the potential of machine learning algorithms for use in healthcare settings and offers valuable insights into the non-invasive identification of anemia using medical imaging. [9]

The literature review explores the use of deep learning (DL) and machine learning (ML) techniques for the diagnosis of anemia, stressing the value of non-invasive approaches for early detection and therapy. It emphasizes how serious a worldwide health concern anemia is, especially iron deficiency anemia (IDA). By examining complex aspects in medical imaging, Automating the diagnosis of anemia requires deep learning (DL) and machine learning (ML). In particular, convolutional neural networks (CNN) are helpful in identifying abnormal blood profiles associated with anemia. The use of ensemble methods combining multiple algorithms also enhances detection accuracy. The methodology followed a systematic approach, utilizing databases like IEEE Xplore, Wiley Online Library, Science Direct, and Google Scholar for pertinent research. Data extraction focused on anemia detection methods, statistical techniques, data sets used, and associated findings, ensuring transparency and ethical considerations. Technologies related to machine learning and deep learning, while not replacing healthcare professionals, are seen as valuable tools in improving diagnostic accuracy, efficiency, and patient care outcomes in healthcare settings, particularly in the non-invasive detection of anemia.[10]

The study by Akinori Mitani and colleagues concentrated on developing prediction models for assessing anemia status based on WHO standard ranges of hemoglobin levels using retinal fundus images from the UK Biobank dataset, which contained 57,163 subjects and a total of 114,326 photos. Three categories were used in the study to classify anemia: approximation anemia, mild anemia, and WHO standard anemia. Researchers looked at a variety of model designs, such as combination models that include both image and metadata elements, age and sex information-based models, and fundus-image-based models. The InceptionV4 architecture was used for modeling, and the Area Under the Curve (AUC) measure was used to assess the performance of the model. The significance of transparent and interpretable deep learning techniques in medical image processing and diagnostics was underscored by the study, which also looked into the interpretability and explainability of the model results.

Many studies have emphasized the advantages and disadvantages of different algorithms in the field of disease prediction using machine learning, as well as the particular applications for which they are most suited. Numerous techniques for supervised machine learning have been investigated by researchers for illness prediction tasks, such as Random Forest, Decision Tree, Naïve Bayes, Artificial Neural Networks (ANN), Vector Machine Support (SVM), Logistic Regression, and K-nearest Neighbor. Notably, Support Vector Machine (SVM) is often used, but Random Forest (RF) has occasionally demonstrated higher accuracy.

Specifically focusing on anemia prediction using Complete Blood Count (CBC) data, researchers have compared the performance of supervised machine learning algorithms such as Naive Bayes, Random Forest, and Decision Tree. The Naive Bayes technique demonstrated higher accuracy compared to Random Forest and Decision Tree in this context [8]. Additionally, innovative approaches involving subsets of classifiers and ensemble learning approaches have been investigated to optimize red blood cell categorization accuracy for the identification of anemia.

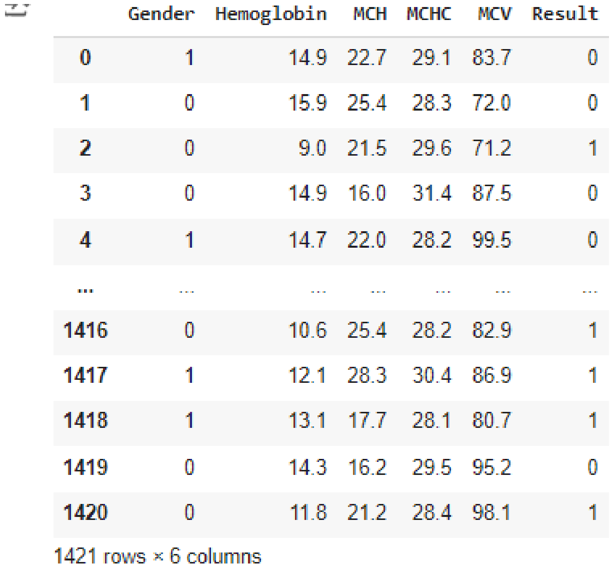
Research has also explored the predictive capacity of common risk variables for anemia status in children under five years of age. In addition to traditional regression techniques, Algorithms for machine learning have shown potential for anemia prediction based on recognized risk variables [11]. Similar to this, machine learning techniques have been used to build prediction models that evaluate the possible risk of anemia in infants and pinpoint important risk variables including exclusive breastfeeding and maternal anemia during pregnancy.

Researchers have discovered that under two-year-old children in Ghana have a greater incidence of anemia when looking at prevalence patterns [14]. Additionally, longitudinal studies have examined how children's anemia status changes over time, emphasizing the ongoing risk of anemia between toddler and preschool age.

III. Methodology

To address our research questions, we proposed a research framework guided by a computational approach. The research process included a literature survey, data collection, algorithm processing, verification, validation, and the presentation of results.

Data Gathering Our original investigation site was Kanti Children's Hospital, where we collected 700 data records of children under 5 years old. From the Complete Blood Count report, we focused on RBC counts to classify anemia. We selected 7 attributes for this purpose: Red Blood Cells (RBC), Hemoglobin (Hb), Hematocrit (HCT), Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), Mean Corpuscular Hemoglobin Concentration (MCHC), and RDW-cv. These attributes were employed to forecast anemia, with



	Gender	Hemoglobin	MCH	MCHC	MCV	Result
0	1	14.9	22.7	29.1	83.7	0
1	0	15.9	25.4	28.3	72.0	0
2	0	9.0	21.5	29.6	71.2	1
3	0	14.9	16.0	31.4	87.5	0
4	1	14.7	22.0	28.2	99.5	0
...	...	...	...	...	...	...
1416	0	10.6	25.4	28.2	82.9	1
1417	1	12.1	28.3	30.4	86.9	1
1418	1	13.1	17.7	28.1	80.7	1
1419	0	14.3	16.2	29.5	95.2	0
1420	0	11.8	21.2	28.4	98.1	1

1421 rows × 6 columns

Fig. 2. Diagonal correlation matrix graph

reference to pediatric standards for children under 5 years, as confirmed by consulting a doctor.

After collecting the data, it underwent preprocessing. The data, recorded in the Hematological Analyzer, was manually collected and then prepared using various preprocessing techniques, including normalization.

Model Preparation Following preprocessing, the dataset was ready for classifier algorithms. We selected six algorithms for anemia prediction: Artificial Neural Network, Support Vector Machine, Naïve Bayes, Random Forest, Decision Tree, and Logistic Regression. We split the data into training and testing sets and used 10-fold cross-validation for validation and verification.

The confusion matrix served as the foundation for performance evaluation, and related formulas were utilized to rate each classifier algorithm's effectiveness. The following performance criteria were also assessed: area under the curve (AUC), recall, F-score, accuracy, and precision. The computation of time-related metrics, such as CPU and wall times needed to execute the algorithms, was also done.

The graph presents a relationship between theoretical quantiles (e.g., 1%, 5%, 10%) on the x-axis, representing percentiles of a normal distribution, and Hemoglobin values on the y-axis. Each plotted point corresponds to a specific Hemoglobin value from the dataset compared against its expected value in a normal distribution at that percentile.

The graph's primary function is to assess the conformity of Hemoglobin values to a normal distribution. A linear arrangement of data points indicates a strong adherence to a normal distribution, while deviations from linearity suggest potential deviations from normality.

This graphical analysis aids in evaluating the suitability of a normal distribution-based deep learning model for predicting anemia based on Hemoglobin data.

Acquiring Data: The collection process involves gathering medical datasets that encompass various blood test results, potentially including parameters like Hemoglobin, Red Blood Cell (RBC) count, Mean Corpuscular Hemoglobin (MCH), and Mean Corpuscular Volume (MCV).

Data Preparation: The absent data points are handled through imputation, and normalization or standardization techniques are applied to ensure uniform scaling across features. Categorical variables may undergo numerical encoding for computational compatibility.

Creation of Models for Deep Learning:

Model Selection: Neural networks with convolutions (CNNs) are frequently used for analyzing microscopic images, such as palm images for anemia identification, depending on the type of data; on the other hand, recurrent neural networks (RNNs) would be a better option for sequential blood test data interpretation.

Model Architecture: CNN architectures typically incorporate convolutional layers for feature extraction, pooling layers for dimensionality reduction, and completely linked layers for classification tasks (e.g., distinguishing anemic from healthy individuals).

Hyperparameter Optimization: Parameters critical to model performance, such as learning rate, layer depth, and neuron count, are fine-tuned to optimize model efficacy.

Training and Evaluation of Models

Data Split: To help with model learning and validation and avoid possible overfitting, the dataset is split into subsets for training and validation.

Training Procedure: Using an optimizer such as Adam, the model undergoes training on the training data, aiming to minimize a defined loss function (e.g., binary cross-entropy for anemia classification).

Performance Metrics: Model performance on the validation set is assessed using metrics like accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC) to gauge classification efficacy.

Additional Considerations

Addressing Class Imbalance: Imbalanced datasets, where anemic cases are underrepresented compared to healthy cases, may require techniques like oversampling or undersampling to mitigate bias and enhance model robustness.

Model Interpretability: Given the inherent complexity of deep learning models, techniques such as Layer-wise Relevance Propagation (LRP) are employed to interpret and discern the influential features contributing to model predictions.

Acquiring Data: The collection process involves gathering medical datasets that encompass various blood test results, potentially including parameters like Hemoglobin, Red Blood Cell (RBC) count, Mean Corpuscular Hemoglobin (MCH), and Mean Corpuscular Volume (MCV).

Data Preparation: Imputation is used to address missing data points. and normalization or standardization techniques are applied to ensure uniform scaling across features. Categorical variables may undergo numerical encoding for computational compatibility.

Creation of Models for Deep Learning

Model Selection:Depending on the type of information, Neural networks with convolutions (CNNs) are commonly chosen for analyzing microscopic images such as palm images for anemia detection, while Neural Networks with Recurrent Architecture (RNNs) may be preferable for sequential blood test data analysis.

Model Architecture:Fully connected layers are frequently utilized for classification tasks (e.g., discriminating between individuals with and without anemia), pooling layers for dimensionality reduction, and convolutional layers for feature extraction in CNN designs.

Hyperparameter Optimization: Parameters critical to model performance, such as learning rate, layer depth, and neuron count, are fine-tuned to optimize model efficacy.

Training and Evaluation of Models: Data Split: The dataset is divided into training and validation subsets to facilitate model learning and validation, preventing potential overfitting.

Training Procedure: Using an optimizer such as Adam, the model undergoes training on the training data, aiming to minimize a defined loss function (e.g., binary cross-entropy for anemia classification).

Performance Metrics: Model performance on the validation set is assessed using metrics like accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC) to gauge classification efficacy.

IV. Conclusion

The significance of the palpable palm in anemia detection, as indicated by previous studies, especially for children, is evident. The novelty lies in utilizing the palm, which outperforms the conjunctiva of the eyes, offering easier access and reduced risk of injury or infection. This study also contributes a novel conceptual framework and dataset published in the Medley repository, advancing the methodology for future anemia detection studies. Moving forward, combining palm, conjunctiva, and fingernail images could enhance anemia detection efficiency.

In this study, we conducted a comparative analysis of CNN, k-NN, Decision Tree, Naïve Bayes, and SVM in detecting anemia using palm images. Our primary dataset consisted of 527 samples, augmented to 2635 using image augmentation to prevent overfitting. The models showed promising results, with Naïve Bayes achieving the highest accuracy of 99.96%, followed closely by k-NN and CNN at 99.92% each, while Decision Tree achieved 97.32%, and SVM reached 94.94% accuracy. Evaluation metrics like recall, precision, F1-score, and AUC were used, validated with tenfold cross-validation and 70/20 split for instruction and assessment.

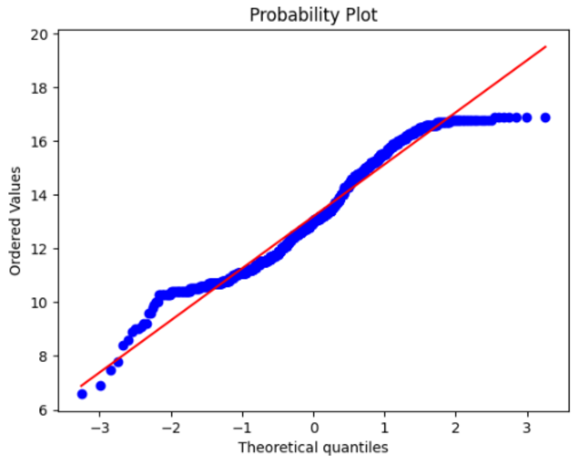


Fig. 3. Performance Analysis of the proposed model

References

[1] Peter Appiahene, Justice Williams Asare, Emmanuel Timmy Donkoh, Giovanni Dimauro, and Rosalia Maglietta. Medical image-based detection of iron deficiency anemia: a comparative analysis of machine learning methods. *BioData Mining*, 16(1):2, 2023.

[2] Marina Barulina, Ivan Ulitin, Tatyana Kaluta, and Alexander Fedonnikov. Using deep learning algorithms for anemia detection: An overview. In Olga Dolinina, Igor Bessmertny, Alexander Brovko, Vladik Kreinovich, Vitaly Pechenkin, Alexey Lvov, and Vadim Zhmud, editors, *Artificial Intelligence in Models, Methods and Applications*, pages 605–615, Cham, 2023. Springer International Publishing.

[3] Maria D Cappellini and Irene Motta. Anemia in clinical practice-definition and classification: does hemoglobin change with aging? *Seminars in Hematology*, 52:261–269, 2015.

[4] Juan P Chalco, Luis Huicho, Consuelo Alamo, Nilton Y Carreazo, and Carlos A Bada. A meta-analysis on the diagnostic accuracy of clinical pallor in children for anemia. *BMC Pediatrics*, 5:46, 2005.

[5] Prakriti Dhakal, Santosh Khanal, and Rabindra Bista. Prediction of anemia using machine learning algorithms. *International Journal of Computer Science and Information Technology*, 15:15–30, 02 2023.

[6] Emerson Meneses, Desislava Boneva, Mark McKenney, and Adel Elkbuli. Massive transfusion protocol in adult trauma population. *American Journal of Emergency Medicine*, 38:2661–2666, 2020.

[7] Akinori Mitani, Abigail Huang, Subhashini Venugopalan, Greg S Corrado, Lily Peng, Dale R Webster, Naama Hammel, Yun Liu, and Avinash V Varadarajan. Detection of anaemia from retinal fundus images via deep learning. *Nature Biomedical Engineering*, 4(1):18–27, 2020.

[8] Dhakal Prakriti, Khanal Santosh, and Bista Rabindra. A comparative study on the use of machine learning algorithms to predict anemia in children under 5 years. *SADI International Journal of Science, Engineering and Technology (SIJSET)*, 9(2):24–37, May 2022.

[9] Susan P Scott, Lenore P Chen-Edinboro, Laura E Caulfield, and Laura E Murray-Kolb. An updated evaluation of the effect of anemia on child mortality. *Nutrients*, 6:5915–5932, 2014.

[10] Tushar N Sheth, Niteesh K Choudhry, Melanie Bowes, and Allan S Detsky. The connection between conjunctival pallor and anemia. *Journal of General Internal Medicine*, 12:102–106, 1997.