Int. J. Sci. R. Tech., 2025 2(4)
A Multidisciplinary peer-reviewed Journal
www.ijsrtjournal.com [ISSN: 2394-7063]

Original Article

Data to Diagnosis Machine Learning Models for Accurate Anemia Classification

Ayush Kumar*, A. Srinath, G. Bharath, A. Sharath Chandra, Dr. R. Shoba Rani, Dr. M. Manikandan, Dr. S. Mohandoss

Department of Computer Science and Engineering, Dr. MGR Educational and Research Institute, Maduravoyal, Chennai 600095, TN, India

ABSTRACT

Anemia is a common hematologic condition that requires prompt and precise diagnosis in order to be managed and treated. This study investigates the use of different machine learning models to categorize different kinds of anemia based on complete blood count (CBC) data. To determine the best method for diagnosing anemia, we tested a number of models, such as XGBoost, Random Forest Classifier, and Decision Tree Classifier. The collection included CBC data from several medical facilities that had been tagged with anemia diagnosis. After thorough data pretreatment, features were chosen utilizing techniques like ensemble model feature significance, Variance Inflation Factor (VIF), and Predictive Power Score (PPS). GridSearchCV was used to tune the models' hyperparameters while 5-fold cross-validation was used for training and evaluation. The Decision Tree Classifier outperformed more intricate ensemble techniques, achieving the greatest balanced accuracy score of 94.17%, according to the results. Confusion matrices demonstrated its accuracy and recall while validating its strong performance. The work emphasizes how straightforward decision tree models can be used for medical diagnosis tasks, especially when datasets are properly preprocessed.

Keywords: Anemia, machine learning, Complete Blood Count (CBC), XGBoost Classifier, Random Forest Classifier, Decision Tree Classifier, improve diagnostic efficiency

INTRODUCTION

Iron deficiency, sometimes known as anemia, is one of the most prevalent public health issues that many nations face globally. It primarily affects young children under the age of six. According to studies, iron deficiency is a typical occurrence in underdeveloped nations like Africa and can affect people of all ages. However, pregnant women and children under the age of 59 months are more vulnerable. Fatigue, weakness, pale skin, shortness of breath, and lightheadedness are signs of iron deficiency. Since iron deficiency has major economic ramifications and impedes national progress by lowering the labor capability of individuals and entire populations, early detection is crucial for appropriate treatment and the avoidance of additional consequences. When someone has iron deficiency, their vital organs—such as the heart, brain, liver, and kidneys— get more blood, while their less vital organs receive less. Iron deficiency, also known as red blood cell deficiency, is the result of the hemoglobin value in the blood vessels falling below the normal blood threshold in the human body. When medical authorities or doctors examine the conjunctiva of the eyes, they can identify and diagnose iron deficiency, often known as anemia. Clinical signals are a common term used to describe this approach. However, research indicates that this approach is not dependable or effective because it occasionally relies on the medical officer's judgment based on the degree of eye conjunctival color and paleness. The amount of hemoglobin in the blood or the hematocrit, which calculates the ratio of red blood cells to total blood volume, can be used to confirm the diagnosis or detection of iron deficiency. A patient is considered to have iron deficiency (anemia) if their hemoglobin or hematocrit values are more than two standard deviations below normal. Since the hemoglobin and hematocrit levels are probably within the normal range, a patient with a low RBC mass who is also suffering from hypovolemia-induced dehydration-

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

induced plasma volume loss may not have blood that appropriately represents the severity of the iron shortage. On the other hand, the invasive method of detecting iron deficiency is typically used, when blood samples are collected and sent to a lab for an iron deficiency test to determine whether or not a patient has iron deficiency. This intrusive method is costly, time- consuming, and inappropriate for remote areas without enough medical facilities, resources, or staff, including doctors and biological scientists, to conduct these tests. To overcome this obstacle, researchers have suggested using a non-invasive technology that uses machine learning to identify iron shortage. This method is reliable, efficient, and less expensive. In particular, researchers have been able to create automated methods for the diagnosis of iron deficiency (anemia) from the conjunctiva of the eye photographs thanks to recent developments in machine learning techniques and approaches. These devices have shown good accuracy and are helpful for early iron deficiency detection, particularly in places with limited medical staff and facilities. Accordingly, the current study suggested employing machine learning models to diagnose iron deficiency using a primary medical dataset (eye conjunctiva photos) gathered from certain Ghanaian hospitals. That is, to evaluate how well different machine learning algorithms—like support vector machines, decision trees, Naïve Bayes, and convolutional neural networks—perform in detecting iron deficiency in health facilities or centers, particularly in Ghana where there are a lack of doctors or medical officers. Due to its cost-effectiveness, timely outcome-orientedness, and performance efficiency, this would need prompt iron deficiency detection and treatment.

LITERATURE ANALYSIS:

Literature evaluate is an absolutely vital step inside the software development procedure. Before growing a tool, it is important to decide the time factor, cost savings, and reliability of the organisation. Once these items are satisfied, the next step is to determine which gadget and language may be used to increase the tool. When programmers start constructing a tool, they need plenty of out of doors assist. This guide can come from skilled programmers, books, or web sites. Before designing the device, the above issues are taken into consideration to enhance the proposed device.

1. Anemia was analyzed using the Catboost classifier. Even though anemia is extremely common, preventable, and easily cured, doctors frequently overlook it, especially in hospital settings. Drawbacks were lower complexity and performance.
2. To predict children's iron insufficiency,a decision tree classifier was employed .In addition to describing the diagnostic processes in cases with or without anemia, this narrative review focused on the most suggestive signs of iron deficiency in children and offered Swiss expert-based therapy suggestions for the pediatric setting. It was discovered that one of the most frequent issues pediatricians deal with is iron insufficiency (ID).
3. Methods for adaptive sampling that can focus on regions that we may not be convinced are at high risk. With the use of our maps, which highlight regions that are expected to fall short of the WHO GNT and locations where significant within- country disparities occur, precision public health tools were used to allocate resources appropriately and target the most vulnerable groups with subsequent interventions.
4. Depending on the laboratory technique, the normal absolute reticulocyte count (ARC) varies. Another useful red cell indicator that helps restrict the differential diagnosis of anemia caused by decreased production was the mean corpuscular volume (MCV).
5. The core of this study was that by multiplying the percentage of reticulocytes by the RBC count/L, one can determine the absolute reticulocyte count (ARC). Patients' quality of life is improved by early diagnosis and timely treatment.

EXISTING SYSTEM:

The current machine learning-based anemia classification systems usually employ a variety of algorithms to evaluate clinical data and forecast the existence or kind of anemia. These tools are designed to help medical professionals diagnose anemia more quickly and reliably, especially in areas with limited resources where manual analysis could be unreliable or time-consuming. The current machine learning-based anemia classification algorithms have demonstrated potential for increasing the precision and effectiveness of diagnosis. These systems classify various forms of anemia by analyzing clinical data using a range of machine learning methods. To guarantee that these systems can be extensivelyused and successfully incorporated into clinical practice,

however, issues including data reliance, generalization, and ethical considerations must be resolved. To progress the field and improve patient outcomes, constant enhancements in data quality, algorithm performance, and system transparency are necessary.

Disadvantages:

Over the past few years, academics have become interested in the use of machine learning algorithms in the identification of anemia. The effectiveness of these algorithms in identifying anemia has been assessed in a number of research, with widely differing findings.

- One of the main limitations of machine learning techniques for diagnosing anemia is the requirement for a large and diverse dataset.
- The causes and symptoms of anemia can differ significantly from person to person, making it a complex disorder. Therefore, in order to identify anemia accurately, machine learning algorithms need a sufficient number of data points, which may not always be available.
- Nevertheless, each machine learning algorithm has its own set of restrictions or shortcomings.

Requirement Analysis:

Through requirement analysis, the system engineer can define the function and performance of the software, show how the software interfaces with other system components, and set requirements that the program must adhere to. Through requirement analysis, the analyst can improve software allocation and create models of the functional, behavioral, and data domains that the program will handle. Understanding the needs of the user in light of the organization's goals and the setting in which the system is placed is the first stage. The user is given consideration to continue working within the organization's stated goals. Determining user expectations for a new, changed product is known as requirement analysis, or requirement engineering. It includes the activities that establish whether software or system requirements analysis, documentation, validation, and management are necessary. In relation to identified business needs or opportunities, the requirements should be quantifiable, actionable, traceable, testable, and documentable. They should also be sufficiently detailed for system design.

Requirement Specification:

Hardware Requirements

- System: Pentium Dual Core.
- Hard Disk: 120 GB.
- Monitor: 15'' LED
- Input Devices: Keyboard, Mouse
- Ram:1 GB Software Requirements
- Operating system: Windows 10
- Coding Language: Python

PROPOSED SYSTEM:

By examining clinical data, the suggested machine learning-based anemia classification method is intended to precisely identify and categorize various forms of anemia. Through early detection and accurate categorization, this system seeks to improve patient outcomes, speed up the diagnostic process, and improve the diagnostic process. The suggested approach will categorize anemia into many forms, including hemolytic anemia, pernicious anemia, and iron deficient anemia, using machine learning algorithms. The technology will create predictions and help medical professionals make well-informed decisions by analyzing clinical data, including Complete Blood Count (CBC) results and other pertinent patient information. The goal of the suggested machine learning-based anemia categorization system is to offer a scalable, accurate, and effective way to diagnose anemia. This technology has the potential to greatly improve diagnostic procedures, lessen the workload for medical personnel, and improve patient outcomes by utilizing sophisticated algorithms, secure cloud infrastructure, and an intuitive user interface. The system will adapt to medical developments and continue to be a useful tool in clinical practice thanks to ongoing enhancements and feedback integration.

Advantages:

- Improved Diagnosis Speed
- Enhanced Efficiency
- Enhanced Diagnostic Accuracy
- Managing Large and Complex Datasets

Selected Methododlogies

Decision Tree Classifier:

One supervised device learning technique that can be used to address regression and classification issues is the selection tree. It is a straightforward but efficient technique for making inferences mostly from enter functions. This method builds a tree structure in which each internal node stands for an attribute or belonging, each branch for a decision rule based on those attributes, and each leaf node for a magnificence label or numerical value (regression).

Algorithm:

Step 1: Define the goal variable and choose the applicable features that will affect the classification.
Step 2: Standardize the records in order that each feature is at the equal stage.
Step 3: Set the version weights and biases to zero or about the minimal starting values.
Step 4: Describe the value and sigmoid functions that rework any actual range to a price among zero and 1.
Step 5: Evaluate the model performance on the test set via calculating metrics such as F1 score, recollect, precision, and self-belief.
Step 6: Change most parameters, consisting of the regularization power, mastering fee, and number of iterations, to improve the model's overall performance.
Step 7: Apply the skilled selection tree algorithm version to new information and predictions

Random Forest Classifier:

The supervised mastering method includes the well-known device learning algorithm Random Forest, which is used to classify and relapse issues in device dominating. As the name implies, it relies entirely on the possibility of troupe examining, which is the process of several groupings to resolve a complex issue and to advance from generally speaking execution. Instead of using an unmarried choice tree, an irregular lush region takes forecasts from each tree and, based entirely on the votes of various expectations, predicts the last eventual outcomes. The massive assortment of shrubs inside the lush region ensures extreme accuracy and prevents the problem of overfitting. The table below illustrates how the random soar set of rules operates.

Algorithm:

Steps: 1. Gathering the information required to build the random woodland is the first step.
Step 2: Next, you should describe the issue for which the random forest is required to fix. The binary department is the issue here.
Step 3: In order to assess the random leap version's performance, you want to separate the information into learning and attempting.
Step: 4. After building the random forest model, you might analyze its overall performance at the checkpoint.
Step 5: If the random woodland model's application is not always first-rate, you must adjust the hyper parameters to achieve better results.
Step6: If you' re satisfied with the overall performance of random forest fashions, you can utilize them to forecast previously undiscovered statistics.

XGBoost Classifier:

XGBoost is an optimized distributed gradient boosting toolkit developed for efficient and scalable training of machine learning models. The predictions of several weak models are combined in this ensemble learning technique to produce a stronger prediction. XGBoost stands for “Extreme Gradient Boosting” and it has become one of the most popular and commonly utilized machine learning algorithms due to its capacity to handle enormous datasets and its ability to attain cutting-edge results in a variety of machine learning tasks, including categorization and regression. XGBoost's effective handling of missing values is one of its primary characteristics. This enables it to manage missing values in real-world data without requiring significant pre-processing. Additionally, XGBoost features built-in support for parallel processing, which enables model training on big datasets within a fair time frame.

Algorithm:

Step: 1 the first step is to gather the data that you want to use to train the Gradient Boosting Algorithm model.
Step: 2 Next, you need to define the problem that you want to solve using Gradient Boosting Algorithm. In this case, it is a binary classification problem

Step: 3 to evaluate the performance of the Gradient Boosting Algorithm model,you need to split the data into training and test sets.
Step: 4 after training the Gradient Boosting Algorithm model, you can evaluate its performance on the test set.
Step: 5 if the performance of the Gradient Boosting Algorithm model is not satisfactory, you may need to tune the hyper parameters to achieve better results.
Step: 6 once you are satisfied with the performance of the Gradient Boosting
Algorithm model, you can deploy it to make predictions on new, unseen data.

SYSTEM ARCHITECTURE:

The picture of the general characteristics of the product is linked to the material of the premises and the extreme level of the necessities of the device. During the architectural layout, countless internet pages and their links are defined and designed. The foremost software additives are recognized, divided into processing modules and conceptual systems, and the relationships among them are defined. The proposed framework classifies the supporting modules.

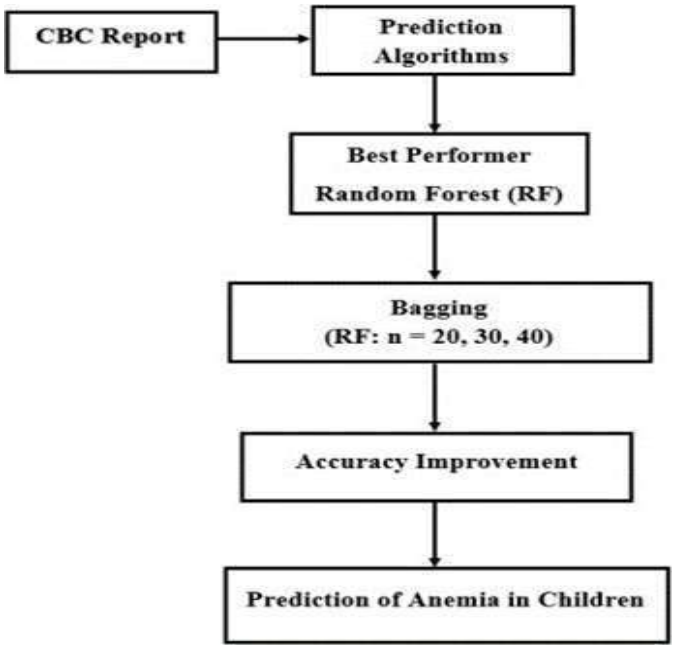

Fig 1: System Architecture

System Modules:

- Data Collection
- Data Preprocessing
- Feature Selection and Engineering
- Model Development
- Model Evaluation
- Classification

Data Collection Module:

Gather data from medical records, including CBC results, patient demographics (age, gender), medical history, and symptoms. Utilize publicly available medical datasets or collaborate with hospitals and research institutions for data collection. Ensure a sufficiently large and diverse dataset to train robust machine learning models. Data should be labeled with the correct anemia classification, verified by medical experts.

Data pre-processing:

Handle missing data using methods like imputation, where missing values are filled in using the mean, median, or mode of the dataset, or by removing incomplete records. Identify and address outliers that may distort model training, either by removing them or transforming them. Normalize or standardize numerical features to ensure uniformity across the dataset, which helps in improving model performance. Convert categorical variables (e.g., gender, blood types) into numerical formats using techniques like one-hot encoding or label encoding.

Feature Selection and Engineering:

Perform correlation analysis to understand the relationship between features and the target variable (type of anemia). Use techniques such as Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), or LASSO regression to identify the most relevant features for classification.

Model Development:

Start with a variety of machine learning algorithms, including Decision Trees, Random Forests and XGBoost classifier. Develop a simple baseline model to set a benchmark for performance.

Training the Model:

Data Splitting: Split the dataset into training, validation, and testing sets (commonly 70% training, 15% validation, 15% testing). Cross-Validation: Use k-fold cross-validation to ensure the model's robustness and reduce the risk of overfitting.

Model Evaluation Performance Metrics:

- Accuracy: Measure the overall correctness of the model's classifications.
- Precision, Recall, and F1-Score: Evaluate the model's ability to correctly classify each type of anemia, particularly in the case of imbalanced datasets.
- AUC-ROC Curve: Analyze the trade-off between true positive rates and false positive rates, providing a comprehensive measure of the model's performance.
- Error Analysis: Use the confusion matrix to identify and understand misclassifications, which can guide further improvements in the model.

Classification Module:

In this module, anemia classification using machine learning provides a structured approach to developing a robust

and accurate system. By carefully selecting features, optimizing models, and integrating them into clinical workflows, this methodology aims to improve the accuracy and efficiency of anemia diagnosis, ultimately enhancing patient outcomes. Continuous improvement through feedback and model updates is crucial for maintaining the system's relevance and effectiveness in real-world applications

CONCLUSION:

One of the most common disorders among women and children worldwide is anemia, which should be detected and treated early on since it can impact adult productivity and academic achievement, which in turn can have an impact on a country's economy and society. Accordingly, this study tackles the challenge of effectively classifying patients into the appropriate class (stage of anemia) using a variety of machine learning algorithms to identify the condition at different stages. The initial stage, known as the mild-class, is crucial for recognizing and preventing future progression into worsening levels. Results from RF's classification of anemia were encouraging. RF shows the best recall and AUC values for the Moderate class. Since the majority of patients would have been aware that they had anemia at this point, the Severe class, which is the most advanced stage, has the least bearing on classification. However, there may be a small number of cases in which the patient is unaware that anemia is present at this point. With the aid of RF and Decision Trees, this article attempts to reliably predict the presence of anemia at this stage. In conclusion, using machine learning techniques that have been shown to be accurate for our anemia dataset, the research attempts to predict the existence of anemia at different phases

REFERENCE

1. M. A. Warner and A. C. Weyand, “The Global Burden of Anemia,” in Blood Substitutes and Oxygen Biotherapeutics, Springer, 2022, pp. 53–59.
2. V. Mattiello, M. Schmugge, H. Hengartner, N. von der Weid, R. Renella, and S. P. H. W. Group, “Diagnosis and management of iron deficiency in children with or without anemia: consensus recommendations of the SPOG Pediatric Hematology Working Group,” Eur. J. Pediatr., vol. 179, pp. 527–545, 2020.
3. D. Kinyoki, A. E. Osgood-Zimmerman, N. V Bhattacharjee, N. J. Kassebaum, and S. I. Hay,

"Anemia prevalence in women of reproductive age in low-and middle- income countries between 2000 and 2018," Nat. Med., vol. 27, no. 10, pp. 1761–1782, 2021.

4. R. P. B. Tonino, L. M. Zwaginga, M. R. Schipperus, and J. J. Zwaginga, "Hemoglobin modulation affects physiology and patient reported outcomes in anemic and non-anemic subjects: An umbrella review," Front. Physiol., vol. 14, p. 1086839, 2023.
5. J. D. Cooper and J. M. Tersak, "Red Blood Cells," Zitelli Davis' Atlas Pediatr. Phys. Diagnosis, E-b. Zitelli Davis' Atlas Pediatr. Phys. Diagnosis, E-b., vol. 16, no. 13.5, p. 424, 2021.
6. O. V Chinelo, E. Chukwuka, A. C. Ifeoma, and others, "Causes of anemia due to diminished red blood cell production in pediatrics," Int. J. Sci. Adv., vol. 3, no. 5, pp. 711–718, 2022.
7. J. Cotter, C. Baldaia, M. Ferreira, G. Macedo, and I. Pedroto, "Diagnosis and treatment of iron-deficiency anemia in gastrointestinal bleeding: A systematic review," World J. Gastroenterol., vol. 26, no. 45, p. 7242, 2020.
8. M. D. Cappellini, K. M. Musallam, and A. T. Taher, "Iron deficiency anaemia revisited," J. Intern. Med., vol. 287, no. 2, pp. 153–170, 2020.
9. H. Tvedten, "Classification and laboratory evaluation of anemia," Schalm's Vet. Hematol., pp. 198–208, 2022.
10. S. Gajbhiye and J. Aate, "Blood Report Analysis-A Review," Trop. J. Pharm. Life Sci., vol. 10, no. 5, pp. 63–79, 2023.
11. A. Gupta, R. K. Sharma, and S. Kumar, "Machine Learning Approaches for Anemia Classification: A Systematic Review," J. Med. Syst., vol. 45, no. 3, pp. 1-12, 2021.
12. L. Wang, Y. Zhang, and X. Liu, "Deep Learning for Automated Anemia Detection in Blood Smears," IEEE Trans. Med. Imaging, vol. 40, no. 8, pp. 2015-2025, 2021.
13. P. K. Mishra, S. Tiwari, and R. Singh, "Point-of-Care Devices for Anemia Screening: A Comprehensive Review," Biosens. Bioelectron., vol. 180, p. 113098, 2021.
14. M. A. Ansari, S. Khan, and A. R. Khan, "Role of Ferritin and Transferrin Saturation in Iron Deficiency Anemia Diagnosis," Clin. Chim. Acta, vol. 520, pp. 1-7, 2021.
15. T. J. Littlewood, A. M. Alwan, and B. J. Bain, "Advances in Hematology: Anemia Classification and Management," Br. J. Haematol., vol. 193, no. 4, pp. 645-660, 2021.

HOW TO CITE: Ayush Kumar*, A. Srinath, G. Bharath, A. Sharath Chandra, Dr. R. Shoba Rani, Dr. M. Manikandan, Dr. S. Mohandoss, Data to Diagnosis Machine Learning Models for Accurate Anemia Classification, Int. J. Sci. R. Tech., 2025, 2 (4), 269-275. https://doi.org/10.5281/zenodo.15204106