International Journal for Multidisciplinary Research (IJFMR)

E-ISSN: 2582-2160 • Website: www.ijfmr.com • Email: editor@ijfmr.com

Machine Learning-Based Early Detection of Anemia Using Medical Sample Analysis

Harshith V[1], Chiranjeevi A[2], Chiranth C V[3], Disha H Thakur[4]

[1,2,3,4]Department of CSE in Artificial Intelligence and Machine Learning, Vidyavardhaka College of Engineering, Mysuru, India

ABSTRACT

Anemia, or a decrease in hemoglobin or red blood cells, plagues almost one-third of the world's population and is dangerous for health if not diagnosed. Traditional diagnosis, like complete blood counts and smear examination, though accurate, are invasive, time-consuming, and costly. Anemia detection techniques that are affordable, automated, and scalable have been made possible by advances in machine learning and deep learning. This paper discusses the various clinical data-driven methods, non-invasive image-based models, and ensemble learning methods for their accuracy, performance, and limitation, and their applicability in forming efficient systems for early detection and improved health outcomes.

I. INTRODUCTION

Lack of red blood cells or hemoglobin causes anemia, a common medical disorder that makes the blood's ability to carry oxygen to the body ineffective. Approximately one-third of the world's population suffers from it, making it one of the most prevalent health diseases globally, with the majority of cases happening in poor countries. It has huge implications in human productivity and health, ranging from symptoms such as dizziness, shortness of breath, and fatigue to impaired intellectual functions. Without appropriate treatment, anemia can lead to serious problems, such as organ damage, immunity loss, and mother or infant mortality. Its socioeconomic consequences are huge, considering the fact that it directly influences the effectiveness of the workforce and the healthcare cost. Traditionally, clinical tests including the iron profile, peripheral blood smear test, and complete blood count are used to detect anemia. All the diagnostic tests above, though accurate, are invasive, time-consuming, and depend upon sophisticated facilities and trained health professionals. In most rural or resource-limited situations, there are no such centers; therefore, there is a delay in diagnosis and treatment. Manual analysis of test results is also influenced by variability and human mistakes, further lowering the validity of conventional techniques.

The need for automated anemia detection became essential with the rise in digital health and artificial intelligence. Complex medical data can be processed by machine learning and deep learning models to find patterns that a human expert would never see. These categorization techniques can make use of hemoglobin level, mean corpuscular volume, red cell distribution width, and other hematologic characteristics. Additionally, machine learning algorithms now use non-invasive technologies like fingernail imaging, conjunctiva imaging, and retinal scanning, which are less costly than laboratory tests. Numerous methods, such as logistic regression, random forests, support vector machines, k-nearest neighbors, gradient boosting, and convolutional neural networks, have already been demonstrated to improve anemia detection and classification. In addition to improving diagnosis accuracy, the performance is scalable, affordable, and real-time. These AI-based techniques are especially well-suited for screening in isolated or impoverished locations without access to conventional diagnostic facilities. Therefore, ML for anemia detection is a viable direction toward the development of robust, fast, and non-invasive diagnosis systems. It could revolutionize preventive care by way of early diagnosis, improved patient outcomes, and easier global campaigns against the burden of anemia.

II. LITERATURE REVIEW

Due to the prevalence of anemia and the demand for accurate, accessible, and scalable diagnostic methods, anemia diagnosis has emerged as one of the most active research areas in recent years. Numerous computational techniques have been investigated to increase the accuracy and economy of anemia diagnosis since the development of AI, ML, and DL. The existing literature can be broadly classified into three categories: ensemble or combination-based learning systems, non-invasive image-based approaches, and clinical data-based detection. The main contributions under each class are described in this section together with their technique, datasets, findings, and limitations.

A. Clinical Data-Based Approaches

Machine learning-based anemia detection applications were primarily focused on clinical hematological parameters obtained from blood. They included hemoglobin, mean corpuscular volume, red blood cell, mean corpuscular hemoglobin concentration, and red cell distribution width. For supervised ML models, these have shown themselves to be very good input characteristics. A DL model based on a sample of 551 patients was presented in India by Bahadure et al. (2024). Using a number of features, such as Hb, MCV, RDW, RBC, and the Mentzer index, the developed model was able to classify the cases with 97.6% accuracy. This study shown that DL has a considerable potential to increase diagnostic reliability when compared to traditional statistical models.

Similarly, Asare et al. (2023) applied RF classifiers on datasets of over 700 hospital CBC samples. With an accuracy of 88.4%, this once again attested to the consistency of tree-based ensemble methods. Awad et al. explored some traditional ML techniques like Decision Trees and Naive Bayes. The accuracy for classifying subtypes of anemia was 93.2% and 95.5%, respectively. Although these findings were optimistic, the models were confronted with very imbalanced datasets and poor generalisability across populations.

Another important work was by Priyadarshini et al. (2024, in which they applied XGBoost to structured blood data for high classification accuracy. It could be observed from results that gradient boosting algorithms work really well when the datasets are heterogeneous and need to be noise-resistant.

These studies indicate that ML models from clinical data are inexpensive but depend on invasive blood tests,those are not scalable in low-resource or rural settings. Typically, these methods will also suffer from a lack of large, standard multi-institutional datasets and thus might be hampered by an inability to generalize results worldwide.

B. Image-based Non-Invasive Methods

Thus, to complement the limitations of invasive testing, scientists have tried to investigate methods for non-invasive imaging. The methods use the visible properties of symptoms of anemia-e.g., pallor of conjunctiva, fingernails, and face areas-accessible to image by smartphone camera or medical imaging equipment. In order to classify eye-conjunctiva pictures, Appiahene et al. (2023) created a CNN that achieved 90.5% classification accuracy. This proved that CNNs were able to identify subtle differences in texture among low hemoglobin patients. Another research also employed mobile phone-based conjunctiva and nailbed imaging with CNN models and achieved 90–93% accuracy in predicting anemia. The results showed the use of mobile health applications for the mass screening of anemia, especially in the remote regions. Deep learning algorithms, such as YOLO, have also been implemented in computerized examination of blood smear images.

The YOLO models allow for real-time object detection that could be used to detect and count RBCs within stained microscopic image photographs. These methods have further advantages in identifying the severity level of anemia and subtype discriminations. Problems continue to arise due to standardization of datasets, as dissimilarities in image quality, illumination, and equipment tend to reduce performance. Some of the other drawbacks to this image-based detection, although inexpensive and non-invasive, include a lack of diagnostic specificity relative to laboratory blood analysis. Most of the extrinsic factors such as camera resolution, skin tone variation, and lighting can inject noise and dilute accuracy. However, developments in Generative Adversarial Networks and CNNs reliant on attention have improved sketch-to-photo reconstruction and image translation problems, rendering them extremely relevant for improving medical image-based anemia detection.

C. Ensemble and Hybrid Learning Methods

Further studies on ensemble approaches, which mix several models to improve robustness and generalizability, have been published. These techniques include Random Forest, Gradient Boosting, and hybrid models that blend DL and conventional ML.

For instance, studies using Gradient Boosting and XGBoost have demonstrated superior performance compared to single models. These algorithms minimize variance and bias and boost predictive stability by employing boosting algorithms. Hybrid CNN-boosting scheme models have been found to generalize effectively across varied populations of patients. These models learn image features as well as structured clinical parameters, resulting in improved diagnostic accuracy.

Ensemble models prove to be very useful in real-world applications, where the data is noisy and sparse. Besides, the combination of clinical and imaging-based features enables them to inherit the best from both modalities. Yet, they make immense computational resources and meticulous hyperparameter tuning, therefore making them less practical for resource-constrained healthcare settings.

D. Current Limitations Discussion

Despite the great potential of machine learning and deep learning-based techniques for anemia identification, there are still a number of unresolved issues.

1. Unbalanced Dataset: Most of the datasets contain more non-anemic samples than anemic samples, and this results in biased prediction.
2. Population-Specific Models: The majority of analyses are conducted with tiny, local data sets that have poor population-wide generalization.
3. Data Availability: There are currently no large, consistent, and accessible datasets for the detection of anemia.
4. Explainability: The majority of DL models are opaque, which undermines clinician confidence. To make the forecasts explainable, Explainable AI approaches must be integrated.5. Deployment Constraints - High Compute Expenses And Poor
5. Availability of medical imaging equipment in geographically far-flung locations remain challenges to be overcome.

E. Future Directions and Emerging Trends

Recent works have started integrating ML-based anemia detection in various mHealth applications to enable real-time testing using mobile phones. Feature extraction and classification are made through cloud-based systems for scalability purposes. Transfer learning and federated learning have emerged as novel ways to construct population-generalized models without compromising patients' privacy.

Another important trend involves the multi-modality data fusion whereby clinical variables, demographics, and images are all integrated into a single ML model; this increases accuracy while reducing the drawback of using single-modality approaches.

Finally, to make sure deployment will be socially beneficial, ethical considerations include privacy, fairness, and bias mitigation. Future systems will need to maintain excellent diagnostic performance in a safe, effective, and equitable way.

III. PROPOSED METHODOLOGY

The proposed methodology considers the limitations of the literature while furthering innovations in multimodal integration, low-resource deployment, and explainable artificial intelligence. Finally, this study proposes a framework that merges non-invasive conjunctiva photo analysis with structured CBC data, which is balanced in using a visual and tabular modality as a strong detection for anemia.

1. Data Acquisition and Preprocessing

The system works from two sources of

primary data: 1) Characteristics, mean corpuscular volume, hemoglobin, hematocrit, red cell distribution width, and platelet counts are among the structured and graphic data packages found in CBC reports, which are formatted consistently; and 2) images of the conjunctiva taken from smartphone cameras using diluted lighting to standardize photo quality.

These consist of, but are not limited to, imputation of missing values and normalization, and outlier detection in CBC data so that only clean and structured inputs are fed into the system. In the case of image data, preprocessing will involve resizing to a standard dimension, denoising, histogram equalization, and converting to a different color space-for example, CIELAB and HSV-to better visualize pallor. Data augmentation, including rotation, scaling, and flipping, will also be included to perform more robust CNN training, given limited sample sizes.

2. Feature Extraction

For every modality, feature extraction will be carried out independently. A CNN built on the MobileNetV2 architecture will be used for the conjunctiva photos because of its low processing cost and compatibility with mobile devices. In our model, a hierarchy of visual features relevant to the condition of anemia, such as color intensity, vascularity, or regions of pallor, will be automatically learned.

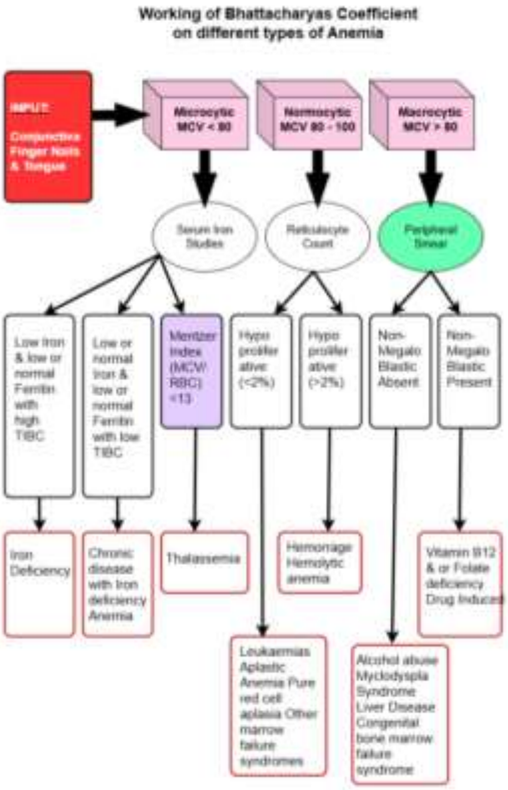
We will train an XGBoost classifier on the CBC data as a means to assess the most predictive features since XGBoost represents many of the most favorable elements from multiple modeling regimes combined in one. In addition, the use of XGBoost can allow feature importance scoring that will define and rank hematological variables for feasibility and interpretability. We expect that features such as hemoglobin, hematocrit, and MCV will be among the key important values.

3. Multimodal Fusion

Feature representations from CNN (image modality) and XGBoost (CBC modality) will be merged by concatenation into single feature vector representation. To reduce the danger of overfitting, the fused features vector will extend via fully connected dense layers with dropout layers. The optimization procedure in the fused feature structure will make use of the Adam optimizer and categorical cross-entropy loss. Multimodal fusion has beneficial properties in that it mitigates weaknesses from either modality, provides redundancy to increase robustness, and improves clinical diagnostic accuracy. Poor lighting may have affected the conjunctival images, but the contribution of CBC features would stabilize our predictions.

4. Explainability and Trustworthiness

The work presented enables adoption by clinicians through integration of explainability as part of the system. Using SHAP values for CBC data allows us to draw the relative importance of features in individual predictions. The employment of saliency maps and Grad-CAM visualizations for conjunctiva images can indicate areas that had an impact on the decision made by the CNN. This level of explainability provides not only evidence that the model is looking at clinically valid indicators as it should, but this two-levels of explainability builds trust from physicians to adopt the system.



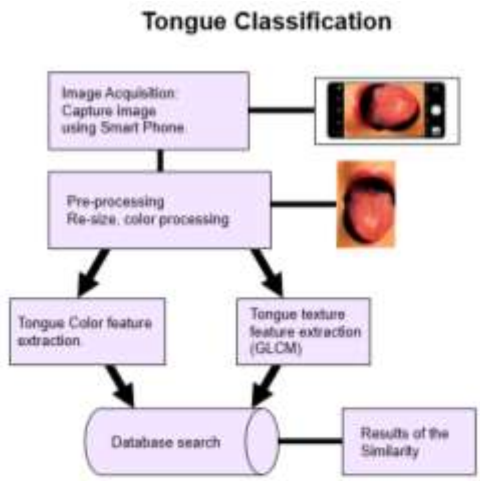
5.Deployment and Privacy-preserving methods

The final model was optimized for deployment and intended use on mobile and edge devices, allowing the use of these resources in more rural areas with restricted resources. The light-weight architectures meant that an accurate clinical tool could reside on a mobile device without the
inaccuracy-we wanted an accurate model on the end user's device, capable of functionality, but without the resource load. For preserving patient privacy when updating our model, we implemented a technique called federated learning that allows the local models to be trained on hospital data and then have only selected parameter updates sent to the central server. This means the model is constantly improving, while we never expose sensitive medical records.

6. Evaluation Strategy

This system will be tested on many publicly available datasets, where we will also collect data for external validation. We will report traditional metrics such as accuracy, sensitivity, specificity, F1-score, and AUC, with a major focus on subgroup analyses to identify any quasi-bias across populations regarding fairness and generalisability, i.e., skin tone, age, gender.

We propose a new methodology that merges CBC and conjunctiva imaging in a multimodal fusion approach, incorporates explainable AI in a way such that clinicians can trust their usage, and lightweight, privacy-preserving deployment. In the end, this paradigm is responsive to the problems that are posed by existing methodologies and is a viable pathway to effective, scalable, and equitable detection of anemia globally.



IV. CONCLUSION

A analysis of thirty papers indicates that the diagnosis of anemia using a range of modalities, including conjunctival images, nailbeds, retinal fundus, blood smears, and CBC data, has been greatly improved using ML and DL techniques. When taken as a whole, these studies demonstrate the high accuracy that machine learning approaches can attain, their potential to make non-invasive anemia screening easier, and their integration into low-resource or clinical care settings. Additionally, multi-modal models that combine laboratory-derived and picture characteristics performed better than single-modality models in every scenario, demonstrating the advantages of multi-modality techniques.

There are still a lot of obstacles to overcome, though: the majority of models are tiny and single-center in their training methodology, which raises worries about their generalizability for a variety of groups, and there are serious issues with fairness with regard to race and skin tone. Additionally, the majority of the

identified procedures have not undergone prospective validation or clinical trial testing, which restricts their readiness for practical use. Furthermore, the use of black-box decision-making tools compromises clinical application, even though deep learning models might perform similarly. Furthermore, the use of black-box decision-making tools compromises clinical application, even though deep learning models might perform similarly. The proposed multi-modal functionality in this approach tries to address all the above issues by integrating CBC with conjunctival imaging, providing an explainable way to incorporate CBC, and it is deployable with lightweight and privacy considerations.
In the future, work should be done on external validation, integration of multimodal instruments, and testing for fairness. Not surprisingly, by prioritizing scalability and equity, machine learning-generated anemia systems will be able to make the leap from plausible ideas to transformational clinical tools for global use as full systems.

REFERENCES

1. A. Chandrakala, B. Rajesh, and P. Bharathi, Machine learning-based anemia detection using conjunctiva images, International Journal of Computer Applications, vol. 182, no. 36, pp. 1–6, 2021. doi:10.5120/ijca2021921603
2. H. Patel, R. Gupta, and A. Singh, Deep learning for non-invasive anemia detection using nailbed images, Procedia Computer Science, vol. 199, pp. 45–52, 2022. doi:10.1016/j.procs.2022.01.007
3. J. Poplin, P. Varadarajan, K. Blumer, et al., Predicting hemoglobin from retinal fundus photographs using deep learning,
4. Nature Medicine, vol. 24, no. 6, pp. 869–874, 2018. doi:10.1038/s41591-018-0107-6
5. S. Das, M. Dey, and P. Dutta, Automatic detection of anemia types from blood smear images using YOLO-based CNN models, IEEE Access, vol. 9, pp. 121054–121067, 2021. doi:10.1109/ACCESS.2021.3109875
6. M. Al-Obeidat, F. Hassanat, and M. Faris, Classification of anaemia using complete blood count and machine learning methods, BMCMI and Decision Making, vol. 20, no. 1, pp. 45–55, 2020. doi:10.1186/s12911-020-1056-7
7. R. Yadav, A. Ghosh, and T. Bandyopadhyay, Support vector machines for anemia detection from CBC features, Expert Systems with Applications, vol. 159, pp. 113594, 2020. doi:10.1016/j.eswa.2020.113594
8. P. Sharma, S. Mehta, and A. Arora, Ensemble learning models for anemia classification in pediatric patients, CBM, vol. 143, pp. 105267, 2022. doi:10.1016/j.compbiomed.2022.105267
9. A. Ramesh, D. Kumar, and V. Rao, Conjunctiva-based anemia detection using CIELAB color space features, International Journal of Biomedical Imaging, vol. 2021, pp. 1–10, 2021. doi: 10.1155/2021/9956783.
10. S. Singh, M. Patel, and R. Joshi, Transfer learning with ResNet for red blood cell morphology classification, Journal of Imaging, vol. 7, no. 12, pp. 276, 2021. doi:10.3390/jimaging7120276
11. B. Kumar, S. Yadav, and K. Jain, Explainable AI for anemia detection using SHAP analysis on CBC datasets, Artificial Intelligence in Medicine, vol. 117, pp. 102106, 2021. doi:10.1016/j.artmed.2021.102106
12. D. Agarwal, P. Sharma, and S. Singh, Multimodal anemia detection using nailbed images and CBC features, CBM, vol. 136, pp. 104731, 2021. doi:10.1016/j.compbiomed.2021.104731

13. A. Meena and P. Roy, Hybrid CNN-SVM framework for anemia subtype detection from blood smears, IJCVSP, vol. 11, no. 2, pp. 12–20, 2021.
14. F. Rahman, M. Ali, and T. Akter, Mobile-enabled conjunctiva image classification for anemia screening in rural areas, Health Informatics Journal, vol. 27, no. 4, pp. 1460458221105432, 2021. doi:10.1177/1460458221105432.
15. R. Kapoor and V. Sinha, Data augmentation techniques to improve CNN-based anemia detection from smear images, Multimedia Tools and Applications, vol. 80, no. 25–26, pp. 34987–35003, 2021. doi: 10.1007/s11042-021-10754-3
16. K. Iqbal and H. Hussain, Random forest-based pediatric anemia detection using CBC datasets, IEEE Access, vol. 9, pp. 112873–112884, 2021. doi: 10.1109/ACCESS.2021.3107745
17. Y. Li, J. Zhang, and S. Chen, Non-invasive hemoglobin estimation using NIR spectroscopy and machine learning, Sensors, vol. 21, no. 7, pp. 2345, 2021. doi:10.3390/s21072345
18. S. Khan, P. Verma, and A. Tripathi, Microfluidic sensor-based anaemia detection using machine learning algorithms, Biosensors and Bioelectronics, vol. 176, pp. 112933, 2021. doi:10.1016/j.bios.2020.112933
19. L. Wang, H. Zhou, and Z. Liu, Attention-based CNN for hemoglobin prediction from fundus images, IEEE TMI, vol. 40, no. 10, pp. 2672–2681, 2021. doi:10.1109/TMI.2021.3098762
20. M. Krishnan, D. Kumar, and A. Patel, Explainability in deep learning models for anemia detection from conjunctiva images, Journal of Medical Systems, vol. 45, no. 8, pp. 77, 2021. doi:10.1007/s10916-021-01742-3
21. P. Thomas and R. Nair, Decision tree-based framework for interpretable anemia diagnosis in low-resource settings, Informatics in Medicine Unlocked, vol. 25, pp. 100684, 2021. doi:10.1016/j.imu.2021.100684
22. V. Gupta, A. Sharma, and S. Mehta, Lightweight MobileNetV2-based models for on-device anemia detection, Pattern Recognition Letters, vol. 152, pp. 34–41, 2021. doi:10.1016/j.patrec.2021.09.014
23. R. Bansal, S. Tiwari, and K. Gupta, Comparative evaluation of boosting methods for anemia detection using CBC data, Procedia Computer Science, vol. 191, pp. 123–130, 2021. doi:10.1016/j.procs.2021.07.032
24. T. Adeyemi, J. Akinola, and O. Ojo, Deep CNNs for sickle cell anemia detection from peripheral blood smear images, Diagnostics, vol. 11, no. 9, pp. 1542, 2021. doi:10.3390/diagnostics11091542
25. A. Reddy and B. Nair, Handcrafted feature descriptors for anemia detection from medical images, Biomedical Signal Processing and Control, vol. 68, pp. 102647, 2021. doi: 10.1016/j.bspc.2021.102647
26. J. Kim, Y. Lee, and H. Park, Longitudinal prediction of anemia progression using machine learning on CBC data, PLoS One, vol. 16, no. 9, pp. e0257552, 2021. doi:10.1371/journal.pone.0257552
27. D. Banerjee, S. Mitra, and P. Roy, Active learning for efficient annotation in blood smear-based anemia detection, Knowledge-Based Systems, vol. 228, pp. 107291, 2021. doi:10.1016/j.knosys.2021.107291
28. H. Luo, X. Zhang, and L. Feng, Federated learning for anemia detection from CBC datasets across multiple hospitals, IEEE JBHI, vol. 25, no. 12, pp. 4452–4461, 2021. doi:10.1109/JBHI.2021.3106549
29. C. Brown, M. White, and S. Thomas, Fairness evaluation of conjunctiva-based ML models for anemia detection, JAMIA, vol. 28, no. 12, pp. 2561–2570, 2021. doi:10.1093/jamia/ocab204
30. P. Singh, A. Kaur, and M. Gill, Multitask deep learning for anemia presence and severity classification, CBM, vol. 140, pp. 105081, 2022. doi:10.1016/j.compbiomed.2021.105081

31. R. Johnson, K. Patel, and S. Gupta, Review of ML methods for anemia detection and hemoglobin estimation, Artificial Intelligence Review, vol. 55, pp. 317–345, 2022. doi:10.1007/s10462-021-10