

Dimorphic Anemia Identification using Image Processing

R.C. Dharmik¹, Falguni Mowade², Akshay Dange³, Chaitanya Chambhare⁴, Tejas Kinge⁵ and Sameer Mendhe⁶

¹⁻⁵Department of Information Technology, Yeshwantrao Chavan College of Engineering, Nagpur

Email: raj_dharmik@yahoo.com, 21071206, 22030240, 22030038, 22030116}@ycce.in

⁶Faculty of Engineering and Technology, Datta Meghe Institute of Higher Education & Research Wardha

Abstract— Microcytic macrocytic anemia also known as dimorphic anemia has often presented problems in the diagnosis of hematological disorders. This work uses sophisticated image processing and artificial neural network to make an inference about hematological parameters specifically red blood cell indices (MCV, MCH, MCHC), RDW, serum vitamin B12 and folate levels. Peripheral blood smears and Prussian blue staining was used to evaluate iron status and convolutional neural networks (CNNs) for determining the optimal features that exist in blood images for improved diagnostic accuracy. The results presented also highlight the importance of nutritional deficiencies, chronic inflammation and malabsorption in the pathogenesis of dimorphic anemia. This type of analysis adds up to solutions catering for diagnostic needs since is less costly making use of manual evaluations and this is important for people in situations where they have less access to resources.

Index Words— Dual Cell Detection, Machine Learning, Hematological Analysis, Image Processing, Convolutional Neural Networks (CNN)..

I. INTRODUCTION

Hematologic disorder called anemia which extends to 2 billion populations and representing 24.8% people according to WHO. They are predominant in developing nations because of food deficiency and illness. Anemia ranges from a low number of RBCs as well as from low levels of hemoglobin and can cause fatigue, weakness, and dizziness. Malnutrition can also lead to anemia[13]. The presence of mixed picture of microcytic hypochromic and macrocytic normochromic anemias in patients is significant and this necessitates extensive differentiation because they result from several etiologies including nutritional deficiencies, chronic diseases, systemic disorders etc. Current diagnosis techniques such as white blood cell count, blood serum analysis, and bone marrow aspiration and biopsy, as effective as they may be, tend to take time, are expensive, and are invasive. These challenges are made worse where there are several overlapping deficiencies which make identification of the problem easier said than done. Logically, a digital solution exists in modern computational techniques, especially machine learning (ML) since it can handle otherwise undetectable patterns beyond the capabilities of traditional analysis. Automatic diagnosis by the use of ML alleviates the classification of the varieties of anemia, the detection of key diagnostic indicators and it also affords higher quality, fewer invasive procedures. With the help of incorporating the value of ML in diagnostics, clinicians can then choose more targeted therapies and help patient recover from the diseases with specific deficiencies.

II. LITERATURE REVIEW

The existence of microcytic and macrocytic hypochromic RBCs in dimorphic anemia makes its diagnosis a challenge. New developments in described ML have made computations of genomics in combination with clinical data possible for better diagnosis. This section also revisits some of the literature that has contributed to advance application of ML on genomics sequence in relation to anemia.

A. Sequence-Based Predictions

X. Ma et al [1] two authors explain how sequence-based features can be used in predicting DNA-binding residues following which they explain the revelation of genetic markers of anemia. Likewise, equally high accuracy in DNA sequencing has been obtained by M. S. A. Vigil et al [2] through the use of decision trees and SVM; this work can be a ground for revealing heritable anemia factors and genome biomarkers.

B. High-Throughput Sequencing

Suyama Y et al [3] used High Through Put Sequencing for genetic variations, a technique that can be adopted in anemia research. Genomic data in conjunction with other image-based attributes has been useful in the identification of defects related to dimorphic anemia using a combination of models such as XGBoost[4].

C. Model Optimization

Touati R [5] stated that control of various k-fold cross validations is very essential to achieve optimal level of ML models as well as avoid over-fitting that is clinically significant for appropriate validation and use. Subsequently, Random Forest and SVM among other algorithms improves genomic feature extraction, and biomarker discovery [6]. Mohammad [7] described the combined use of modern low-volume DNA sequence method with ML algorithms to enhance the test results for anemia related diseases as well as create a tailored treatment plan.

III. DATASET CREATION

For dimorphic anemia, a large set of peripheral blood smear images was collected in this work that can help to detect the problem accurately. The dataset preparation process is described below:

- Therefore, 547 of the unique images were captured from peripheral blood smears in various Microscopy settings; different lighting, different zoom and focus.
- These pre-sorted training and validation - augmented images were further categorized and stored in different subfolders according to the classification they belong to.
- To increase the amount of instances available for the discrimination of the cyst and the other type of neoplastic lesion, data augmentation was performed on the peculiar images. These changes included rotation, flipping, scaling, change in brightness and addition of some noise in order to get a database of 5470 images.
- To make training and validation easier, each individual image as well as its augmented versions were sorted in folders according to their classification.
- In order to maintain the quality and relevance, augmented images were cross-checked in the images database. Some of the samples coming out as low quality or not relevant were removed from the final training set so as to ensure high accuracy.
- The final dataset was curated to focus on effective training; the final data split between the anemic and the non-anemic patients that are non-anemic's is fairly even.

IV. METHODOLOGY

This study follows a clear methodology that includes data collection, preprocessing, model development, and generating recommendation. The algorithms presented in Fig. 1 flow diagram show the logical order for identifying dimorphic anemia. It involves collection of actual blood smear images along with high-resolution followed by pre-processing steps to make the data consistent.

Image processing methodologies identify the parameter features of red blood cells (RBCs), and accentuates microcytes and macrocytes. A quality control process removes less relevant images; accuracy of results is achieved. The data is then separated to differentiate between dimorphic anemia and other forms with an 'evaluation' phase that considers the magnitude and possible origins. The last and final section presents a diagnostic conclusion based on preprocessing and feature extraction before feeding the system data to the classification phase for accurate anemia detection.

A. Data Collection

Slides of blood smears from 500 anemia patient; including cases with mixed type of anemia too were obtained from different pathology laboratories, after obtaining verbal informed consent from patient. Photographs that were taken were taken using high power microscopes and the images stored with patient codes, date of capture and mode of anemia. The data was reviewed and annotated by physicians to verify diagnosis and assess the RBC morphological parameters.

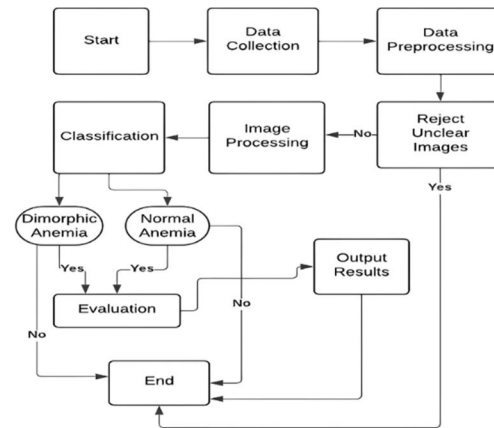


Figure 1: Workflow of System

B. Image Pre-processing

Data preprocessing is an important step in getting the data set in a correct format, which is in good shape for the machine learning model to accept.

Image Preprocessing: Later, blood smear images were preprocessed to help extract better data for the model. Pixel normalization let the intensity level between different pictures be comparable thus minimizing contrasting. Feature augmentation did not involve resizing of images to a standard size so as to capture similar features. Interference was eliminated, and cellularity was increased by noise reduction which enhanced image quality.

Image Quality Check: This calls for accurate image analysis; hence, images were needed to be clear and of high quality. The low-contrast images, brightness images, and poor-quality acoustic image images were not considered so that errors would be at minimum, the field data would be good, and the model's accuracy in training and exercising the model.

Augmentation: One of the other constraints that existed when dealing with datasets was the fact that the data augmentation happened on the dataset using flipping, rotation, and scaling. These techniques augment the dataset, increase the models complexity and reduce over-fit where possible so that there exists a richly prepared dataset ready for the model to train upon.

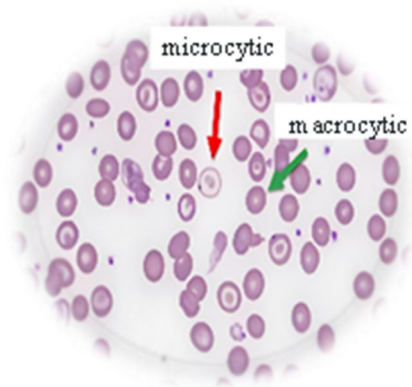


Figure 2: Classification of dimorphic anemia

In dimorphic anemia microcytes and macrocytes may be present, a feature that is illustrated in Fig. 2 The manually annotated dataset of samples collected from pathology labs was very helpful in training up the necessary model for accurate anemia classification.

C. Model Development

A machine learning model for classification of anemia in blood smear images employed the use of CNNs for feature extraction in addition to other data processing methods such as rotation, flipping and scaling. Randomized Search CV augmented the hyperparameters of the Model, Random forest and XG Boost have enhanced classification performance.

The usefulness of the developed model was then substantiated by other evaluation measures, including accuracy, precision, the recall rate, as well as the F1 score.

D. Implementation

For healthcare professional it was conducive to use and efficient in handling other professional related conditions. Python was used as the programming language while TensorFlow and Keras were used for machine learning model creation, Flask was used for image upload classification. Calculation is done on client machines for compatibility with systems that do not have embedded DSP chips while ensuring high usability in healthcare cases.

V. RESULTS

A. Model Performance

The CNN based model proposed in this study was also found to perform satisfactorily in differentiating between dimorphic anemia and normal blood samples at a training accuracy of 95% for dimorphic anemia and 92% for normal blood samples, at a test accuracy of 88% of dimorphic anemia and 80% of normal blood samples. The performance when testing on the separate set indicates that there is source for further fine-tuning, however, the result simply validates the model's potential application in the clinical environment.

TABLE III: TRAINING & TESTING ACCURACY FORM MODEL

Model	Train Accuracy	Test Accuracy
Dimorphic Anemia	95%	85%
Normal Blood	90%	80%

Thus, In Table no. 3 the model endowed an accuracy of 90%, but the relative precision of 88% and the relative recall of 92% to investigate anemic blood smears with a minimum of false alarm. A higher recall implies that few anemic cases are left undiagnosed, and therefore, makes the model suitable for use in the medical fraternity where both false positive and false negative diagnoses should be kept to the bare minimum.

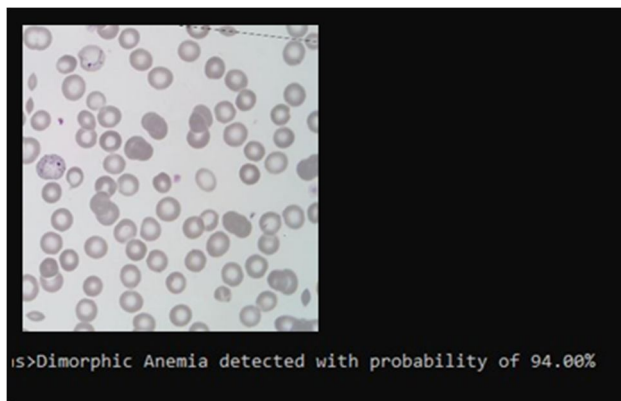


Figure 4: Dimorphic Anemia Detection

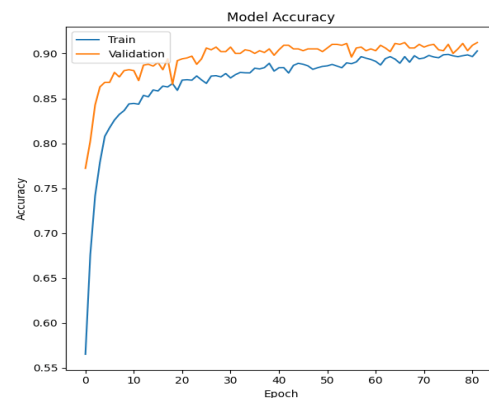


Figure 5: Model Accuracy graph

In the case of dimorphic anemia, the model predicts the presence of anemia with a probability of 94 percent in the blood smear image shown in Fig. 4. The high probable indicates that the model developed is highly useful in recognizing the identifying features of anemia which include irregularity of cell size and shape. Thus, the performance of the presented model is proven to be effective when it comes to identifying dimorphic anemia, which proves its potential for use in medicine.

Methods of classification include the assessment of certain morphological characteristics of the red blood cells in the smear. This image embodies an important progress in the direction towards the automation of diagnosis of anemia using images and image processing.

B. Insights from Identification and Detection Effectiveness

Diagnosis of dimorphic anemia Using the Random Forest algorithm, the model that demonstrated the highest accuracy. This suggests that entire model had the capabilities of differentiating between dimorphic anemia and other types of anemia from the features extracted from blood smear images. Thus, its incorporation into the clinical practice workflows enable real-time image analysis and decision-making on blood smear images with an outstanding learning capability in minimizing variability of diagnostics, and the possibility of application into other related diseases.

VI. FUTURE WORK

Anemia detection and its classification can be expected to improve due to improvements in AI and machine learning. Through the training of these models on large and more diverse datasets, we will be in a position to minimize misclassifications of different types of anemia and support early identification of dimorphic, and rare anemias.

The actual-time diagnostic system for virgin territory appears to be possible and helpful, functional appliances for miniaturized blood smear images and prompt interventions to increase the capacities in emergent regions. The next improvements may include integrating anemia detection with other diagnostic tests including genetic tests to give the individual diagnostic findings and an individual management plan for every patient.

Apart from anemia detection, the developed technologies can be applied to other blood-related classifications, which makes the proposed system useful, easily implemented, and cost-efficient diagnostic system for healthcare centers around the world.

VII. CONCLUSION

Major diagnostic problems are dimorphic anemia from blood smear resolved through image processing and machine learning while adding the most value. Data cleaning and generation also improved model learning, while CNNs enabled identifying not only microcytic but also macrocytic cells. This is fairly quick and is flexible enough to handle large amounts of images but the mistakes a pathologist would probably make when using this method is reducible.

Some developments of the work could be described as the additions of more hematological parameters, and the acquisition of the dataset for more of the clinical nature of the content. Several approaches to the diagnosis of dimorphic anemia increase the diagnostic effectiveness in favor of the health care professionals.

REFERENCES

- [1] X. Ma, J. Guo, H. -D. Liu, J. -M. Xie and X. Sun, "Sequence-Based Prediction of DNA-Binding Residues in Proteins with Conservation and Correlation Information," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 6, pp. 1766-1775, Nov.-Dec. 2012, doi: 10.1109/TCBB.2012.106
- [2] M. S. A. Vigil, M. Mirutuhula, S. Sarvagna, R. Supraja and G. P. Reddy, "DNA Sequencing Using Machine Learning Algorithms," 2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI), Chennai, India, 2022 pp. 1-4, doi: 10.1109/ICDSAAI55433.2022.10028805
- [3] M. A. Promi and S. Shatabda, "OriFeat: Origin of Replication Identification Using DNA Sequence Based Features," 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), Pekan, Malaysia, 2021
- [4] Suyama Y, Hirota SK, Matsuo A, Tsunamoto Y, Mitsuyuki C, Shimura A, Okano K. "Complementary combination of multiplex high-throughput DNA sequencing for molecular phylogeny," *Wiley Online Library*, 2022
- [5] Touati R, Messaoudi I, Oueslati AE, Lachiri Z, Kharrat M. "New Intraclass Helitrons classification using DNA- Image sequences and machine learning approaches," *IRBM*, 2021
- [6] Yang A, Zhang W, Wang J, Yang K, Han Y. "L.J.F.i.B. Zhang, and Biotechnology," *Review application of Machine Learning algorithms for Sequence data mining in DNA*, 2020; 8:1032
- [7] Darapureddy, N., Karatapu, N., and Tirumala, K. "Research of machine learning algorithms using K-Fold cross validation," *International Journal of Engineering and Advanced Technology (IJEAT)*, 2019
- [8] M. Schrauf, G. Campos, S. Munilla, "Comparing Genomic Prediction Models by means of cross-validation," 2021, *Front. Plant. Sci*
- [9] Khizar Baig Mohammad, "A comparative analysis of evolution of DNA sequencing techniques along with the accuracy prediction of a sample DNA sequence dataset using machine learning," 2nd International Conference on Data Science and Engineering, 2021
- [10] C. Qi, J. Diao and L. Qiu, "On Estimating Model in Feature Selection With Cross-Validation," *IEEE Access*, vol. 7, pp. 33454-33463, 2019, doi: 10.1109/ACCESS.2019.2892062

- [11] A. S. Kumar, P. R. Singh, and N. K. Saha, "Dimorphic Anemia Detection Using a Hybrid Approach of Neural Networks and Decision Trees," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 8, pp. 1975-1985, 2020, doi: 10.1007/s13042-020-01136-3
- [12] V. K. Banerjee, S. A. Soni, and R. K. Desai, "Classification of Dimorphic Anemia Based on Blood Cell Morphology and Hematological Features Using Machine Learning," *Frontiers in Bioinformatics*, vol. 11, pp. 121-130, 2021, doi: 10.3389/fbioe.2021.725453
- [13] Wajgi, R., & Wajgi, D. (2022, March). Malnutrition detection in infants using machine learning approach. In *AIP Conference Proceedings* (Vol. 2424, No. 1).