Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 12, 2025

THALASSEMIA CARRIER IDENTIFICATION FROM BLOOD MICROSCOPYIMAGES: A COMPARATIVE ANALYSIS OF ML AND DL APPROACHES

Muhammad Hammad[1], Joddat Fatima[*2] Nadia Sultan[3]

[1,*2]*Department of Software Engineering, Bahria University, Islamabad, Pakistan*
[3]*Department of Electrical Engineering, Bahria University, Islamabad, Pakistan*
[*2,3] *Centre of Excellence in Artificial Intelligence (CoE-AI), Bahria University, Islamabad, Pakistan*

[*2]joddat.fatima@bahria.edu.pk

DOI: https://doi.org/10.5281/zenodo.18058674

Keywords
Thalassemia carrier screening, public health in Pakistan, Healthcare accessibility, AI-assisted diagnosis, Blood smear microscopy, Machine learning

Article History
Received: 27 October 2025
Accepted: 11 December 2025
Published: 26 December 2025

Corresponding Author: *
Joddat Fatima

Abstract

Thalassemia is one of the most prevalent genetic disorders globally, especially in regions like Asia and among individuals of African descent. In Pakistan, the carrier rate for thalassemia is between 5% and 7%, affecting over 10 million people, with approximately 5,000 children born annually with β-thalassemia major (βTM). Early detection of carriers is vital for managing and preventing severe cases. This study introduces a machine learning approach to detect thalassemia carriers through blood smear image analysis. We preprocessed the images to extract features such as color, texture, and shape from a dataset of 7,108 blood images, representing nine cell types related to thalassemia. Various machine learning and deep learning models were applied to classify the images as thalassemic or non-thalassemic. Among the machine learning models, Random Forest achieved the highest accuracy at 91.1%, while MobileNetV2 led among deep learning models with a 90% accuracy. These promising results suggest potential for real-world application in Thalassemia screening programs, contributing to improved diagnostic methods and healthcare outcomes for affected populations.

INTRODUCTION

Thalassemia is a group of blood disorders caused by abnormal hemoglobin production, result-ing in fewer red blood cells. It is a common genetic disorder worldwide, particularly prevalent in populations of Mediterranean, Middle Eastern, Southeast Asian, and African descent (Za-heer et al., 2020). It poses a significant healthcare burden worldwide, particularly in countries like Pakistan, which has a high prevalence of thalassemia carriers. The two standard types of thalassemia are alpha thalassemia and beta thalassemia. When the production of alpha globin chains is disturbed, alpha thalassemia occurs. When the production of beta-globin chains is disturbed, beta thalassemia occurs. The severity of thalassemia symptoms can vary widely, rang-ing from mild and manageable to severe and life-threatening. Treatment options include regular blood transfusions, iron chelation therapy, and, bone marrow transplants, which are expensive and available in only three medical facilities in Pakistan (Muncie & Campbell, 2009). Genetic counseling is also advised for thalassemia-

affected individuals and their families to comprehend inheritance patterns and make informed decisions.

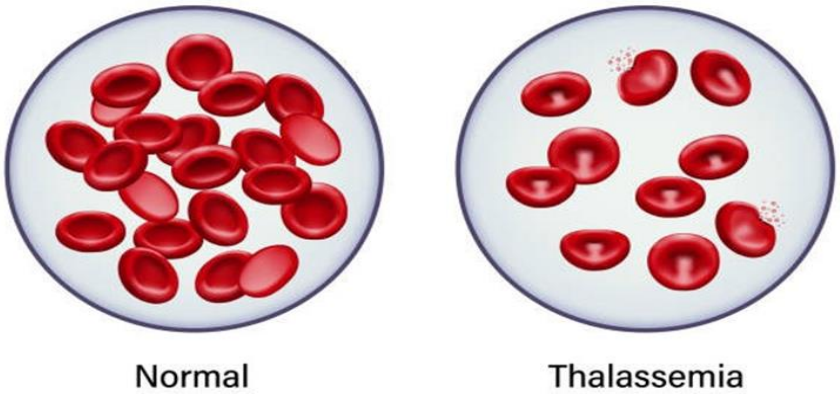

Fig. 1 Normal Vs. Thalassemic Red Blood Cells (Wanda, 2025)

Normal blood cells exhibit various characteristics. They are typically of regular size and shape and appear as biconcave discs. In thalassemia cells, hemoglobin synthesis is impaired, leading to various abnormalities in red blood cells size, shape, and function. Specific characteristics can be observed in the blood cells of individuals with thalassemia cells.

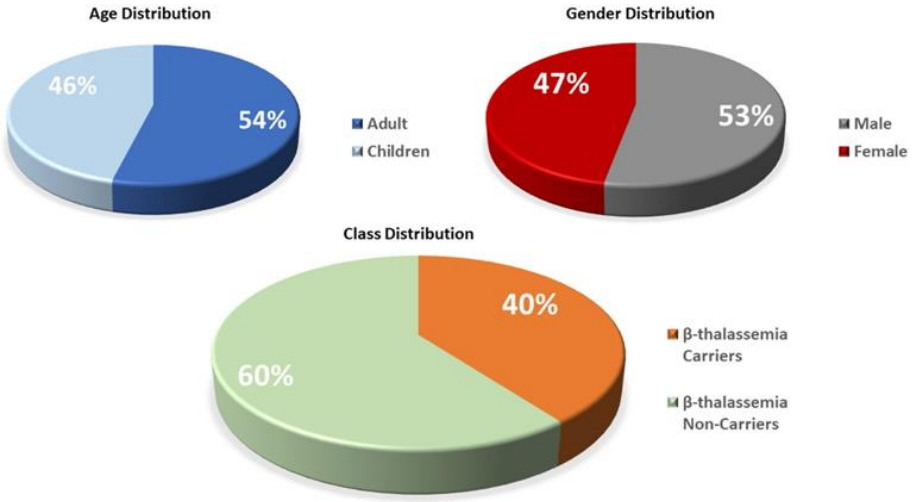

Fig. 2 Thalassemia Statistical Data in Pakistan (Rustam et al., 2022)

The estimated population of Pakistan is approximately 225,633,392 (225 million), and the frequency of the β-thalassemia (β -thal) ranges between 5.0% to 7.0%, indicating the presence of more than 10 million carriers in the country. Moreover, the annual incidence of β-thal major (β-TM) in Pakistan is around 5000 children (Khaliq, 2022). The management of Thalassemia remains challenging due to the lack of standardized protocols and the heavy reliance on blood transfusions (Zaheer et al., 2020).

The complicated healthcare delivery system in Pakistan, which both federal and provincial governments run, makes the problem even more difficult to handle. This fragmentation creates challenges in implementing cohesive strategies for early detection, prevention, and treatment of thalassemia. Moreover, the socioeconomic landscape of Pakistan poses additional hurdles in

the management of thalassemia. The majority of the population belongs to lower socioeconomic strata, making it financially burdensome for families to afford the long-term treatment and management required for thalassemic children. Prompt identification of thalassemia carrier traits allows for the implementation of appropriate management strategies, resulting in improved patient outcomes and quality of life. Early detection is critical for treatment of severe forms of thalassemia because it allows for timely initiation of treatments, including regular transfusions of blood, chelation therapy to eliminate excess iron and potentially curative bone marrow transplants (Shikha et al., 2021). Accurate and early detection of thalassemia carriers using blood images has significant implications for improving patient care, enabling timely interventions, and facilitating appropriate management strategies for individuals affected by this genetic blood disorder. Our study aimed to utilize optimal machine learning algorithms to detect and classify thalassemia from blood smear images. Thalassemia carrier identification provides a diagnosis and facilitates timely intervention for the affected individuals.

Related Work

The literature on thalassemia diagnosis highlights several approaches and methodologies. Some researchers have integrated clinical reports with blood smear images for detection, employing clinical tools and deep learning algorithms for feature extraction, followed by classification using machine learning models. Additionally, studies have explored image segmentation with deep learning architectures such as U-Net, along with transfer learning techniques and data engineering methods to enhance accuracy. Despite these successes, ongoing research continues to explore emerging trends and future directions in medical image processing (Ker et al., 2017).

(Shikha et al., 2021) proposed a diagnostic approach that combined clinical reports with blood smear images to improve the identification of thalassemia, aiming to overcome the challenges in its diagnosis. Clinical features were extracted using a blood analyzer, whereas a deep convolutional neural network (CNN) was employed to extract visual features. These features are then fused, and principal component analysis (PCA) is applied to reduce the computational cost and eliminate redundancy, thereby generating an optimized feature set. Machine learning algorithms with high sensitivity, specificity, and accuracy, such as Naive Bayes, Random Forest (RF), and KNN are employed for classification. This approach achieves up to 81.5% accuracy in classifying thalassemia image features.

(Ahmad et al., 2018) proposed a methodology combining image preprocessing techniques, erythrocyte segmentation, and extraction of various morphological features, including Fourier descriptors, Hu moments, Zernike moments, and geometrical features. The study utilized digitized blood smear images from healthy individuals, as well as from patients with iron deficiency anemia (IDA) and thalassemia, captured under a light microscope. Two separate experiments were conducted: one analyzing 24 morphological features and the other analyzing 31. Statistical methods were used to assess and compare the features. Logistic regression yielded the best results, with optimal feature subsets achieving 83.5% accuracy, 83.5% sensitivity, and a positive predictive value of 83.3%. The study concluded that this methodology effectively classified abnormal and normal erythrocytes in IDA and thalassemia.

In (Sharma et al., 2016) focused on detecting sickle cell anemia and thalassemia, two common genetic disorders affecting approximately 3.2 million people globally. They proposed a method that involves capturing thin blood smear images and preprocessing them using a median filter. Overlapping erythrocytes were segmented through marker-controlled watershed segmentation, followed by the application of morphological techniques to enhance image quality. Key attributes such as aspect ratio, radial signature, metric value, and variance were extracted from the images. To classify three specific erythrocyte morphologies, elliptocytes, dacrocytes, and sickle cells associated with sickle cell anemia and Thalassemia, a K-nearest neighbor classifier was trained using 100 images. This algorithm improves the speed, effectiveness, and efficiency of both the

training and testing processes, achieving an accuracy of 80.6% and a sensitivity of 87.6%. In (Sadiq et al., 2021) proposed using red blood cell indices from a complete blood count (CBC) test as a quick and affordable method for screening individuals. This approach is more expensive, time-consuming, and requires equipment-intensive high-performance liquid chromatography (HPLC) tests. This study introduces an ensemble model called SGR-VC, which combines Random Forest, Gradient Boosting Machine (GBM), and Support Vector Machine (SVM) algorithms. With an accuracy of 93%, their comparative analysis demonstrates the SGR-VC model's effectiveness in screening for β-Thalassemia carriers. A summary of research employing machine learning (ML) approaches to identify Thalassemia from blood smear images is shown in the table below.

Table 1. Summary of ML Models from literature

#	Author	Techniques	Dataset	Results
1	Shikha et al., 2021	Random Forest & CNN	Image features combined with CBC test report	81.5%
2	Ahmad et al., 2018	Logistic Regression	Real Blood Smear Dataset from Hospital Canselori Tuanku Muhriz (HCTM).	83.5%
3	Sharma et al., 2016	K-Nearest Neighbors	Images of blood smears infected with sickle cells, dacrocytes, and elliptocytes	80.6%
4	Sadiq et al., 2021	Random Forest, SVM	Punjab Thalassemia Prevention Project Lab Reports dataset	93%

In (Amira et al., 2022) developed and evaluated a supervised semantic image segmentation model using the U-net architecture. This approach incorporates data engineering techniques such as annotation, augmentation, preprocessing, and transfer learning. To improve accuracy, Prediction Time Augmentation (PTA) was applied. Results showed a mean Intersection Over Union (IoU) score of 88% with PTA and 82% without PTA. An inverse relationship was found between the combined loss score and thalassemia prediction. The qualitative analysis revealed that the model accurately identified codocytes (target cells), and streamlined the final images compared with the original annotated ground truth.

In (Khan et al., 2022) proposed a novel method for automated thalassemia assessment using hemoglobin (Hb) electrophoresis images, to support hematologists in resource-limited settings. This study used a dataset of 524 Hb electrophoresis images from 824 subjects. The methodology has two main components: (1) lane extraction-based segmentation of electrophoresis images, and (2) binary classification (normal or abnormal) using deep convolutional neural networks (CNNs) and transfer learning. Among the various models tested, InceptionV3 and MobileNetV2 performed the best, achieving accuracies of 95.8% and 95.72%, respectively. MobileNetV2, being a lightweight model, offers low latency and is well-suited for mobile applications. This approach provides high accuracy and is expected to enable rapid and reliable thalassemia detection using Hb electrophoresis images.

Paper (Ali et al., 2023) proposed a segmentation method that combines pre-treatment procedures and image processing techniques. The study explored 11 color spaces, six filters, three contrast enhancement methods, and used fuzzy c-means and K-means algorithms. The segmentation performance was evaluated using five metrics based on ground truth images. The focus was on automatic red blood cell (RBC) segmentation from microscopic blood smears, specifically analyzing how thalassemia affects the RBC shape. Ground truth images were created using Photoshop for multi-object detection (RBCs). Local datasets from Thalassemia patients and healthy individuals were used to optimize the process, with images captured under various lighting conditions, with and

without a yellow filter. The best results accuracy of 0.91±0.14 and performance of 95.34% were obtained under medium light without the yellow filter.

In study by (Rodellar et al., 2018) proposed a method for automatic blood cell recognition, focusing on malignant lymphoid and blast cells. This process involves segmentation, feature extraction for the nucleus and cytoplasm, and classification using supervised machine learning. The segmentation achieved 98.9% efficiency, reducing 2464 initial descriptors to 150 key features. An SVM classifier was employed, achieving 90.3% accuracy in distinguishing the seven groups of abnormal lymphoid cells and normal lymphocytes. This approach effectively combines segmentation, feature extraction, and classification to identify specific blood cell types.

In the literature reviewed, deep learning (DL) models are predominantly utilized for tasks such as image segmentation and augmentation in the context of thalassemia detection. Research that uses deep learning (DL) approaches to identify β −thalassemia from blood smear images is compiled in the table. 4.

Table 2. Summary of DL Models from literature

#	Author	Techniques	Dataset	Results
1	Amira et al., 2022	U-net architecture	500 images	88%
2	Khan et al., 2022	InceptionV3, MobileNetV2, ResNet50	Hb electrophoresis image dataset	95%
3	Ali et al., 2023	Fuzzy c-means, K-means algorithms	Blood cells images (258)	0.91±0.14%
4	Rodellar et al., 2018	SVM classifier	Abnormal lymphoid cells (AL)	90.3%

Based on the literature review, gaps in the existing research on thalassemia detection using blood smear images include the limited use of deep learning and machine learning, lack of standardization in methodologies and datasets, insufficient focus on specific cell types, inadequate validation and generalization, and a lack of explainability in deep learning and machine learning models. To address these gaps, this study proposes a comprehensive approach that integrates deep learning and machine learning techniques, establishes a standardized methodology, focuses on specific cell types, and conducts extensive validation. Our contributions aim to advance the field by improving the accuracy and efficiency of thalassemia detection, ultimately leading to more effective diagnostic tools for this genetic disorder.

Materials and Methodology

Overview

Our research process involved several key steps. First, red blood cell images were collected and preprocessed to prepare them for model training. The data were then divided into training and testing sets. Subsequently, both machine learning (ML) and deep learning (DL) models were applied, and the performance was evaluated based on accuracy. We utilized two available datasets of blood smear images specifically related to thalassemia, which are as follows:

Dataset

The dataset (Tyas et al., 2022) includes 12,500 augmented images of blood cells (JPEG) with cell type labels (CSV). Each of the four cell types had approximately 3,000 images, which were organized into four different folders. Cell types include eosinophils, lymphocytes, monocytes, and neutrophils. This dataset was complemented by a supplementary dataset comprising the original 410 images (pre-augmentation). The dataset is publicly available at Kaggle.

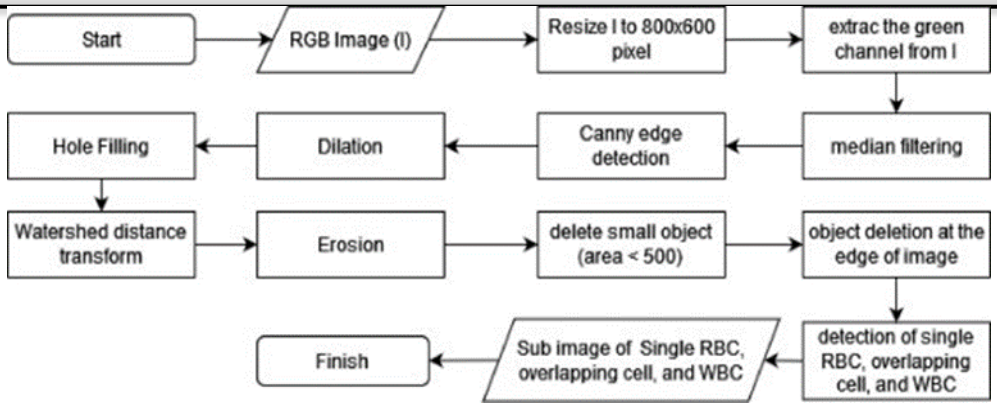

Fig.4 Abstract Diagram for Dataset (Tyas et al., 2022)

Erythrocyte (red blood cell) dataset in Thalassemia case

The data collection was comprised of 7,108 images of single red blood cells, representing nine distinct cell types, all with a resolution of 800 × 600 pixels. Microscopic images were captured from peripheral blood smears using an Olympus CX21 microscope and Optilab Advance Plus camera. During preprocessing, the original images, which measured 4100 × 3075 pixels (RGB), were resized to 800 × 600 pixels to reduce computational complexity. The data were formatted in PNG for grayscale images.

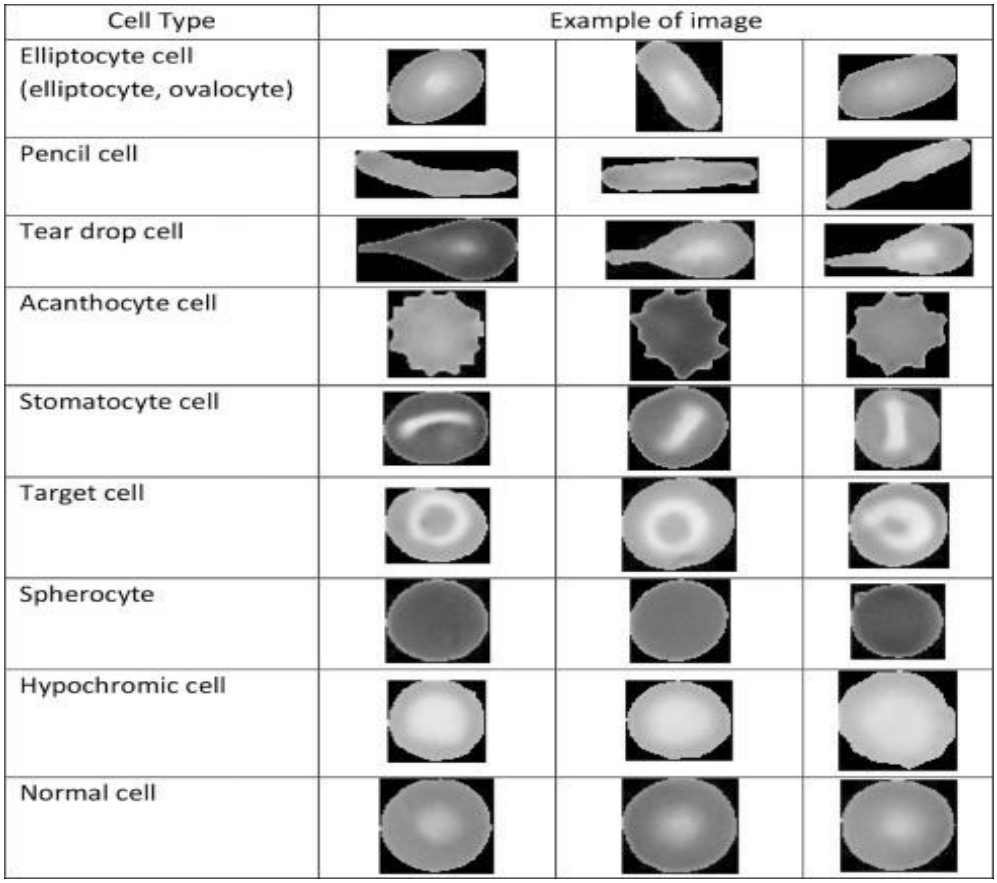

Fig. 3 Nine types of cells in Thalassemia case with their shapes (Tyas et al., 2022)

This study focuses on preprocessed erythrocyte images specifically related to thalassemia. Notably, this dataset was not used in previously published studies. The preprocessing steps enhanced the images for further analysis. This unique dataset serves as a valuable resource for exploring thalassemia through machine learning and deep learning techniques, facilitating insights into the classification and detection of thalassemia from blood smear images.

Proposed Methodology

The proposed system automatically and accurately classifies whether a patient is healthy or is affected by thalassemia. Initially, red blood cell images were uploaded for preprocessing, which helped to identify their corresponding classes. After preprocessing, the thalassemic images were uploaded to extract features, followed by the upload of non-thalassemic grayscale images. The architecture of the proposed model is as follows:

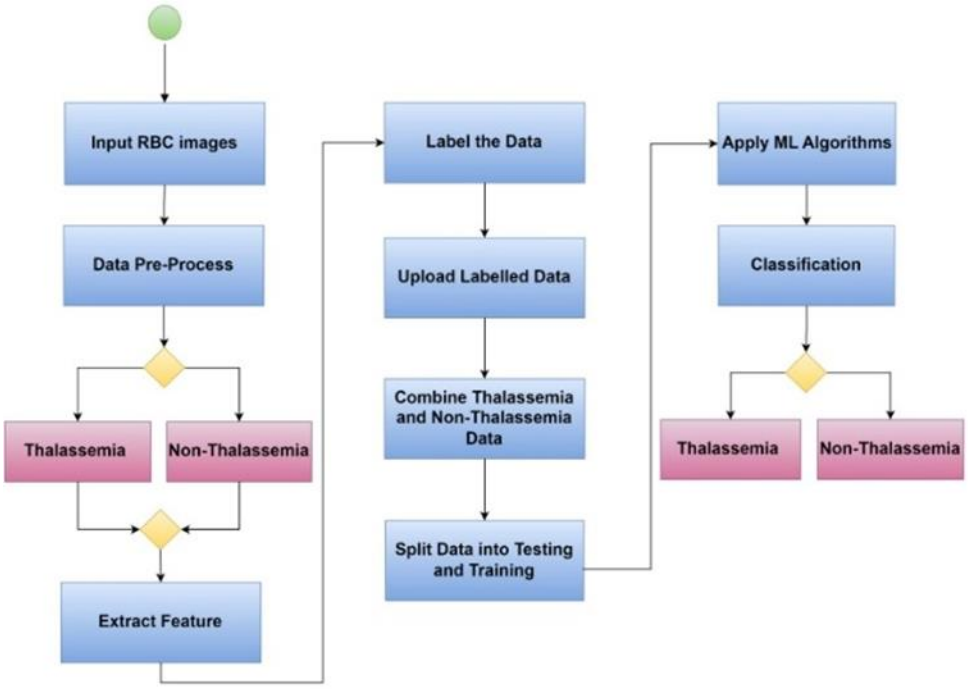

Fig.5 Proposed System methodology which takes RBC images as an input, preprocesses the images to extract features, and applies ML models to classify whether the input is a thalassemia carrier or normal.

The system accepts an image of red blood cells as the input, serving as the primary data source for analyzing and classifying thalassemia carriers. The proposed model consists of three steps: preprocessing, feature extraction, and classification. Each step is explained in detail below:

Preprocessing

The preprocessing steps for the red blood cell images in this study included several crucial stages to prepare the images for segmentation and analysis. First, the images were resized from 4100 × 3075 pixels to 800 × 600 pixels to reduce the computational load in the later phases. Subsequently, the green channel of the RGB images was isolated for further processing. Several image enhancement techniques have been applied, including median filtering, Canny edge detection, dilation, and hole filling, to improve the image quality and enhance the visibility of red blood cells. Additionally, after applying a Watershed (WDT) algorithm to separate overlapping erythrocytes, the images underwent erosion, and small objects with an area less than 500 pixels were removed. Finally, the incomplete cell shapes at the edges of the images were discarded from the dataset. These preprocessing steps are essential for ensuring the accuracy and reliability of the segmentation

process, which is critical for the subsequent analysis and classification of the red blood cells (Tyas et al., 2022). These preprocessing steps are annotated in the dataset.

Feature Extraction

First, the Thalassemic images were uploaded to extract the features simultaneously, followed by the non-thalassemic grayscale images. From the input image, the following three feature categories were extracted:

Color Features

The system calculates the mean and standard deviation of the grayscale image, which are the key statistical measures of pixel intensities. The mean represents the average pixel intensity, indicating the overall brightness of the image, whereas the standard deviation reflects the spread or variation of pixel intensities around the mean. A higher standard deviation indicates greater contrast, meaning that the pixel values are more widely dispersed. These measurements provide insights into the image brightness and contrast, which are essential for tasks such as feature extraction and image enhancement.

Texture Features

The Grey Level Co-occurrence Matrix (GLCM) is used to extract texture features (GLCM). GLCM describes the frequency of various pixel intensity combinations in an image. From the GLCM, the code computes four texture features: contrast, dissimilarity, homogeneity energy. These features provide information about the texture patterns in the image.

Shape Features

Shape features are extracted by calculating the contour area of objects in a binary image, where contours represent the boundaries of objects, and provide valuable information regarding their size and shape. Before this, the original image is converted into a binary format, where the pixels are either black (0) or white (1), representing the background and foreground, respectively. By calculating the contour areas, the system extracts shape features that are useful for tasks such as object recognition, classification, and analysis.

Once the features are extracted, a feature set is created for both Thalassemic and non-Thalassemic cases, resulting in two labeled CSV files. Machine learning (ML) models were then applied to improve the efficiency and accuracy of the classification process. The models were trained on 80% of the data, and the remaining 20% was used for testing. To ensure consistent data partitioning each time the code was executed, the random state was set to 42. The random_state parameter ensures that the data are split in the same manner every time the code is run. Without a fixed random state, the data can be shuffled differently each time, leading to slightly different training and testing sets. Setting random_state=42 (or any other constant value) ensures consistency across runs, making the results reproducible. By fixing the random_state, we can run the same code and obtain identical results. 42 is often used as a default or an example value, but it can be any number. The choice of 42 is arbitrary, made famous by popular culture (Douglas Adams' The Hitchhiker's Guide to the Galaxy where 42 is "the answer to life, the universe, and everything").

Limitations Due to Preprocessed Dataset

The dataset used in this study was publicly available and had undergone prior preprocessing steps, including resizing to 800 × 600 pixels, contrast enhancement, and segmentation to isolate individual red blood cells. While these steps were essential for our analysis, they limited our ability to perform ablation studies to evaluate the impact of each individual preprocessing technique on model performance. Without access to the raw, unprocessed images, we couldn't systematically assess how each preprocessing component influenced the results. In future work, we plan to utilize raw blood smear images to conduct comprehensive ablation studies on individual preprocessing steps. This approach will help us gain deeper insights into the necessity and effectiveness of each technique, ultimately refining our methodology and improving model accuracy.

Classification Techniques

ML Model

Logistic regression, well-suited for binary classification, achieved a precision of approximately 0.856 and a recall of 0.862, indicating that it correctly predicted 85.6% and 86.2% of positive instances, respectively. The Decision Tree model, which is effective for complex data, showed a precision of approximately 0.863 and a recall of 0.859, with 86.3% and 85.9% correct predictions of positive cases, respectively. SVM, ideal for high-dimensional data, reached a precision of 0.834 and recall of 0.843, predicting 83.4% and 84.3% of positive instances respectively. The CNN model demonstrated a precision of 0.875, recall of 0.878, and an F1 score of 0.867, with an accuracy of 87.8%. Random Forest performed best overall, with a precision, recall, F1 score, and accuracy of 90.9%, 91%, 91%, and 91.1%, respectively.

Table 3 Comparison of Different Machine Learning Models

Number	Parameters	Logistic Regression	Decision Tree	SVM	CNN	Random Forest
1	Precision	0.856	0.863	0.834	0.875	**0.909**
2	Recall	0.862	0.859	0.844	0.878	**0.911**
3	F1-score	0.858	0.861	0.837	0.867	**0.910**
4	Accuracy	86.2%	85.9%	84.3%	87.8%	**91.1%**

Random Forest Model

The Random Forest ensemble learning technique uses several decision trees to increase the accuracy of the classification or (regression). It produces a forest of trees, each trained using a random feature set and portion of the data.

Random Forest performs well when data are imbalanced or highly complex. The model achieved a precision of approximately 0.91, indicating that it correctly predicted the positive class 91% of the time. The sensitivity (or recall) was approximately 0.911, indicating that the model correctly identified 91.1% of actual positive instances.

Table 4 Random Forest model results in Precision 90.9%, Recall 91.0%, F1 score 90.9% and accuracy 91.1%

Number	Parameters	Score
1	Precision	0.909174
2	Recall	0.910689
3	F1 Score	0.909751
4	Accuracy	0.910689

The F1 score, which is the harmonic mean of precision and recall, was approximately 0.91. Overall, the model achieved an accuracy of 0.911, indicating that it correctly classified 91.1% of the test set instances.

DL Models

We also implemented several deep learning (DL) models to compare their performances with the machine learning (ML) models. The DL models VGG16, ResNet50, MobileNetV2, EfficientNetB4, and DenseNet121 were used. MobileNetV2 achieved the highest accuracy of 90%, followed by DenseNet121 at 84%. The ResNet50 model achieved an accuracy of approximately 72%, whereas EfficientNetB4 achieved the lowest accuracy at 49%.

Table 5 Comparison of Different Machine Learning and Deep Learning Models

Model	Type	Accuracy Score
VGG16	DL	69%
ResNet50	DL	71.6%
MobileNetV2	**DL**	**90%**
EfficientNetB4	DL	49%
DenseNet121	DL	84%

When comparing the results of machine learning (ML) and deep learning (DL) models, MobileNetV2 achieved the highest accuracy (90%) among the DL models. In contrast, Random Forest outperformed all models, with the highest accuracy of 91.1% among the ML models. The CNN model performed well, with an accuracy of 87.8%. However, the EfficientNetB4 model lagged behind, with a significantly lower accuracy of 49%, indicating poor performance compared with the other models.

Table 6 Comparison of the proposed model with previous models

Author	Model/Technique	Parameter (Accuracy)
Shikha Purwar, et al.	Random Forest	81.5% accuracy, 0.85 AUC in ROC
Izyani Ahmad, et al.	Logistic Regression	83.5% accuracy, sensitivity, and positive predictive value
Our Proposed System	**Random Forest, Logistic Regression**	**91.1% accuracy with the RF model and 86.2% accuracy with Logistic Regression**

The table above highlights that the proposed system enhances the accuracy in both Random Forest (RF) and Logistic Regression models. The RF model proposed by (Shikha et al., 2021) work achieved a maximum accuracy of 85%, whereas that of the proposed system reached 91.1%. Similarly, (Ahmad et al., 2018) obtained a Logistic Regression model with an accuracy of 83.5% accuracy, whereas the proposed system improved the accuracy to 86.2%. Although the comparison is based on the model performance, the datasets are different, as the dataset used in the proposed system has not been previously used in any published research.

Statistical Analysis

K-Fold Cross-Validation is a statistical method used to assess the performance of ML models by dividing the dataset into K subsets. K = 5 was used, meaning the dataset was split into 5 parts. The model is trained on 4 out of the 5 folds and tested on the remaining fold. This process was repeated 5 times, with each fold being used as the test set once. This reduces model overfitting or underfitting by ensuring that every data point is used for both training and testing.

A Wilcoxon Signed-Rank Test was performed to compare the accuracy scores of two models. We selected Logistic Regression and Random Forest. The p-value obtained from the test was 0.4375. A p-value of 0.4375 is greater than the common significance level of 0.05, indicating no statistically significant difference between the performance of both models on this dataset. Similarly, Friedman Test was performed to compare the accuracy scores of three models that are Logistic Regression, Random Forest, and MobileNetV2. The p-value obtained from the test was 0.2466. A p-value of 0.2466 is greater than 0.05, indicating no significant difference in the performance of the three models. This suggests that Logistic Regression, Random Forest, and MobileNetV2 perform similarly with respect to accuracy.

Discussion and Conclusion

The study provides an in-depth exploration of thalassemia detection through blood smear images using machine learning (ML) and deep learning (DL) approaches, bridging critical gaps in prior research. By employing a comprehensive methodology that involved preprocessing erythrocyte images, extracting diverse features (texture, color, and shape), and applying multiple classification techniques, this research demonstrated notable advancements in thalassemia diagnosis. Among ML models, the Random Forest algorithm achieved the highest accuracy (91.1%), showcasing its robustness in handling complex datasets and imbalanced classes. This performance was supported by high precision, recall, and F1 scores, affirming the model's reliability. Compared to existing studies, such as those by Purwar et al. (81.5% accuracy) and Ahmad et al. (83.5% accuracy), the proposed system outperformed in accuracy, indicating the efficacy of the tailored preprocessing and feature extraction techniques. In DL models, MobileNetV2 emerged as the most effective, achieving 90% accuracy, followed by DenseNet121 (84%). These results reflect the potential of lightweight and efficient architectures like MobileNetV2 in handling medical imaging tasks, particularly when resources are constrained.

This study successfully classified thalassemia carriers and non-carriers using blood smear images, achieving the highest accuracy among ML models with Random Forest (91.1%) and among DL models with MobileNetV2 (90%). These findings underscore the significance of comprehensive preprocessing and feature extraction in improving model performance. The proposed system sets a benchmark for thalassemia detection, with applications in early diagnosis, public health interventions, and resource-limited settings.

Future Work

Building on the promising results of our current study, we recognize that expanding the dataset is crucial for improving the generalization and robustness of our models. In future research, we plan to gather a more diverse local dataset by collaborating with multiple medical facilities. This expanded dataset will include raw blood smear images from patients of different ages, genders, ethnic backgrounds, and geographical regions. By incorporating a wider variety of morphological variations associated with thalassemia, we aim to enhance the ability of our models to generalize across different populations and imaging conditions. Working with raw, unprocessed images will also enable us to perform comprehensive ablation studies on individual preprocessing steps. This will allow us to systematically evaluate the impact of each preprocessing technique on model performance, providing deeper insights into which steps are most critical for accurate classification. Such analysis is essential for refining our methodology and improving the overall robustness of the models. Through these initiatives, we aspire to enhance the robustness, accuracy, and clinical applicability of our proposed models. Expanding the dataset and refining our methodology will ultimately contribute to more effective thalassemia screening programs, facilitating early diagnosis and intervention for affected individuals.

Conflict of Interests

There is no conflict of interest to report.

Data Availability Statement

The dataset is available on Kaggle.

REFERENCES

Ahmad, I., Sheikh Abdullah, S. N. H., & Raja Sabudin, R. Z. A. (2018). Morphological features analysis for erythrocyte classification in IDA and thalassemia. *International Journal of Advanced Computer Science and Applications, 9*(12). https://doi.org/10.14569/IJACSA.2018.091253

Ali, N. J., Yaba, S. P., & Sardar, P. (2023). Automated thalassemia cell image segmentation using hybrid Fuzzy C-Means and K-Means. *Zanco Journal of Pure and Applied Sciences, 35*(4), 22–33. https://doi.org/10.21271/ZJPAS.35.4.3

Amira, J. Z., Makki, M., & Kassem, R. (2022). Thalassemia diagnosis through medical imaging: A new artificial intelligence-based framework. In *2022 International Conference on Smart Systems and Power Management (IC2SPM)* (pp. 41–46). IEEE. https://doi.org/10.1109/IC2SPM55349.2022.9798584

Khan, M. S., Ullah, A., Khan, K. N., Riaz, H., Yousafzai, Y. M., Rahman, T., Chowdhury, M. E. H., & Abul Kashem, S. B. (2022). Deep learning assisted automated assessment of thalassaemia from haemoglobin electrophoresis images. *Diagnostics, 12*(10), 2405. https://doi.org/10.3390/diagnostics12102405

Khaliq, S. (2022). Thalassemia in Pakistan. *Hemoglobin, 46*(1), 12–14. https://doi.org/10.1080/03630269.2022.2026261

Ker, J., Wang, L., Rao, J., & Lim, T. (2017). Deep learning applications in medical image analysis. *IEEE Access, 6*, 9375–9389. https://doi.org/10.1109/ACCESS.2017.2788044

Muncie, H. L., Jr., & Campbell, J. S. (2009). Alpha and beta thalassemia. *American Family Physician, 80*(4), 339–344.

Rodellar, J., Alférez, S., Acevedo, A., Molina, A., & Merino, A. (2018). Image processing and machine learning in the morphological analysis of blood cells. *International Journal of Laboratory Hematology, 40*(Suppl 1), 46–53. https://doi.org/10.1111/ijlh.12816

Rustam, F., Ashraf, I., Jabbar, S., Tutusaus, K., Mazas, C., Pascual Barrera, A. E., & de la Torre Díez, I. (2022). Prediction of β-thalassemia carriers using complete blood count features. *Scientific Reports, 12*(1), 19999. https://doi.org/10.1038/s41598-022-23812-y

Sadiq, S., Khalid, M. U., Ullah, S., Aslam, W., Mehmood, A., Choi, G. S., & On, B. W. (2021). Classification of β-thalassemia carriers from red blood cell indices using ensemble classifier. *IEEE Access, 9*, 45528–45538. https://doi.org/10.1109/ACCESS.2021.3066851

Sharma, V., Rathore, A., & Vyas, G. (2016). Detection of sickle cell anaemia and thalassaemia causing abnormalities in thin smear of human blood sample using image processing. In *2016 International Conference on Inventive Computation Technologies (ICICT)* (Vol. 3, pp. 1–5). IEEE. https://doi.org/10.1109/INVENTIVE.2016.7830076

Shikha, P., Tripathi, R., Ranjan, R., & Saxena, R. (2021). Classification of thalassemia patients using a fusion of deep image and clinical features. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 410–415). IEEE. https://doi.org/10.1109/CONFLUENCE51648.2021.9377166

Tyas, D. A., Ratnaningsih, T., Harjoko, A., & Hartati, S. (2022). Erythrocyte (red blood cell) dataset in thalassemia case. *Data in Brief, 41*, 107886. https://doi.org/10.1016/j.dib.2022.107886

Wanda. (2025, July 15). *Malaria – world map*. Instituut voor Tropische Geneeskunde. https://www.wanda.be/en/a-z-index/malaria-world-map/

Zaheer, H. A., Waheed, U., Abdella, Y. E., & Konings, F. (2020). Thalassemia in Pakistan: A forward-looking solution to a serious health issue. *Global Journal of Transfusion Medicine, 5*(1), 108–110. https://doi.org/10.4103/GJTM.GJTM_6_20