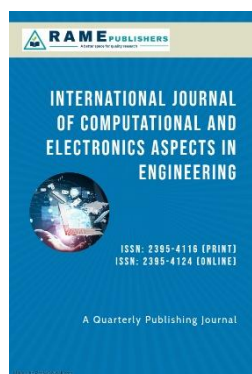


A Secure AI-Based Thalassemia Diagnosis Framework

Razan Abdulhammed* 

Artificial Research Center, Northern Technical University, Mosul, Nineveh Governorate, 41000, Iraq

*Correspondence: rabdulhammed@ntu.edu.iq



Abstract—Thalassemia is an illness that affects about ten million human beings worldwide, most notably in poorer nations. To improve the results of patients suffering from the illness, prompt diagnosis of thalassemia is of utmost importance. Unfortunately, many traditional diagnostic techniques suffer from delays in screening, lack of specialists, and ethical issues related to data protection. Artificial intelligence can help to devise reliable methods for diagnosis of thalassemia in adults and children. The present study proposes an accurate and efficient diagnostic model based on artificial intelligence for the diagnosis of thalassemia using machine learning and deep learning techniques. AES-256-GCM method is used to secure sensitive medical information, thus ensuring its confidentiality and integrity throughout storage and transfer processes. The proposed framework integrates data acquisition, preprocessing, and diagnostic decision support. Naive Bayes, Tabular Transformer, Random Forest, and RNN models are trained and evaluated using the PTPP dataset. Model performance is assessed using accuracy, precision, recall, and F1-score. The Random Forest outperforms other classifiers with an Accuracy of %99.7. The proposed secure AI diagnostic framework has the potential to support healthcare professionals by enabling earlier, more consistent, and data-driven screening for Thalassemia. By combining intelligent diagnosis with encryption-based data protection, the framework may contribute to secure, scalable, and privacy-aware medical screening technologies.

Keywords: Thalassemia Diagnosis, Machine Learning, AES-256-GCM Encryption, Random Forest, Medical Data Privacy

Article – Peer Reviewed

Received: July 10, 2026

Accepted: August 31, 2026

Published: September 14, 2026

Copyright: © 2026 RAME Publishers

This is an open access article under the CC BY 4.0 International License.



<https://creativecommons.org/licenses/by/4.0/>

Cite this article: Razan Abdulhammed, “A Secure AI-Based Thalassemia Diagnosis Framework”, *International Journal of Computational and Electronic Aspects in Engineering*, RAME Publishers, vol. 7, issue 4, pp. 389-398, 2026.

<https://doi.org/10.26706/ijceae.7.4.20260807>

1. Introduction

Thalassemia is an inherited blood disorder passed from parents to their biological children through genetic mutations or deletions in the genes responsible for hemoglobin production in the human body. Here, the body will not be able to produce enough hemoglobin as the patient gets older in age. This deficiency in hemoglobin production can lead to anemia. The symptoms of Thalassemia include Jaundice, fatigue, bone changes, and growth delays [1]. Thalassemia symptoms range from mild to severe, depending on the severity of the patient's hemoglobin deficiency. Based on the affected hemoglobin gene chain, doctors classify Thalassemia into two types : alpha and beta Thalassemia [2]. Thalassemia can be diagnosed through various medical procedures, such as blood tests, genetic testing, and hemoglobin analysis. Many people of different ages suffer from this inherited disorder. There are between 300K to 400K patients with Thalassemia [3][4]. Furthermore, 50K to 60K births are registered as new severe cases per year. Moreover, approximately 80 million to 350 million people carry the disorder but are not infected with it [1]. This means that there is a global carrier prevalence of almost 5 % of the world population [1]. To be specific, Thalassemia affects people with ancestry from the Mediterranean, Middle Eastern, African, and South Asian regions [5]. The only cure for this disease is a stem cell transplant. Yet, genetic counseling can help families understand the risk of passing it on to children. This disease can affect the life of the infected person. However, patients with mild cases can have a normal life expectancy. While patients with severe cases require lifelong treatment, if this treatment is continuously supervised and monitored by a healthcare provider, they can still achieve good long-term outcomes. With recent developments in technology, AI, and healthcare services, developers and healthcare clinicians have attempted to provide affordable, low-cost solutions and applications to diagnose Thalassemia through the CBC blood Test. Different studies with different objectives have been conducted by researchers, developers, and healthcare providers.

The next section consolidates previous engineering solutions and experimental attempts relevant to the current study. Also, the section details the model, datasets, and evaluation strategies employed earlier. The main contributions of this study are summarized as follows:

1. A systematic literature review is conducted to identify open-source datasets relevant to the problem domain and to highlight AI techniques that have not yet been applied to such datasets by prior researchers.
2. A systematic comparative analysis of 4 classifiers is carried out.
3. An AI Framework is proposed to diagnose Thalassemia.
4. Integrates a data encryption mechanism to protect sensitive patient information against unauthorized access and potential breaches.

The rest of this paper is organized as follows: Section II glance at a systematic literature review of relevant work. Section III lists and identifies publicly available datasets of Thalassemia. Section IV explains the proposed system design. The classifier models and decision equations are introduced in Section V. The results and Discussion are presented in Section VII. Finally, the Conclusion is reported in Section VIII

2. Systematic literature review of relevant work

This section presents a synthesized overview of the systematic literature review conducted before this study. The purpose of this review is to determine and assess the methods of AI applied in the diagnosis of thalassemia, as well as to point out where the said methods have not yet been examined. The review was organized in accordance with the previously established research questions described further down.

RQ1: Which machine learning and deep learning algorithms have been used for Thalassemia diagnosis?

RQ2: How effective are machine learning and deep learning models in classifying Thalassemia types using publicly available datasets?

RQ3: What publicly available datasets and features are utilized for training machine learning and deep learning models to achieve optimal classification performance of types of Thalassemia?

RQ4: What emerging technologies are not yet incorporated into existing research?

RQ5: What datasets are not yet incorporated into existing research?

was According to the predefined research questions, a set of studies were explored during the systematic literature review process and phases. Eligibility, whether to include the study or not, was determined using a structured set of inclusion and exclusion criteria, which were applied consistently across all retrieved studies during both the abstract and full-text screening. The study arranges these inclusion and Exclusion criteria in an enumerated list as shown below, starting with the inclusion criteria.

First, the review includes every study that matches the following set of inclusion criteria:

1. The study utilized a publicly available dataset.
2. The study focuses specifically on the classification of Thalassemia types.
3. The study was published between 2021 and 2026
4. The study provides specific performance metrics such as Accuracy and Precision.
5. The study uses structured datasets stored in CSV, TSV, Excel, or other tabular formats.

Second, the study excludes any study that matches one of the following sets of exclusion criteria.

1. The study focuses on general blood disorders or non-specific anemia.
2. The study relies solely on private or proprietary clinical datasets that are not accessible to the public.
3. The study used Image-Based Datasets such as PNG, JPG, and DICOM.
4. The study used Unstructured Data such as Text, audio, video, or graph-only datasets.

As mentioned earlier, this section focusses on reporting studies that utilize the PTPP publicly available dataset [6]. To start, Authors in [7] evaluated XGBoost, SVM, and KNN on PTPP dataset [6], The XGBoost classifier achieved accuracy %99.5 while SVM classifier achieved accuracy of %99.4. Rustem et al. [8] evaluate different machine learning and deep learning methods on PTPP dataset [6], as highlighted in Table II, among the ML classifier RF, and Extra Tree classifier achieved an accuracy of %96.0 in which the dataset was oversampled using SMOTE and integrated with a unified framework of PCA and SVD. Whereas Gradient Boosting Machine (GBM) showed the lowest accuracy of %89.0.

Table 1. studies that utilize PTPP Dataset

Study	Models Used on PTPP
M. U. Nasir et al. (2025) [7]	XGBoost, CNN
M. U. Nasir et al. (2025) [9]	KNN, SVM, XGBoost
Shoaib et al. (2026) [10]	FL with Explainable AI
Rustam et al. (2022) [8]	CNN, PCA, RF, LR, ADA, SVC, ECT, LSTM, GRU, CNN-LSTM

Publicly Available Datasets

This section reports the publicly available datasets identified during the systematic literature review. This paper includes datasets related to blood tests since the concrete aim of this study is to develop an AI diagnosis system able to diagnose thalassemia based on blood tests. A key finding from the systematic review is that most datasets are private clinical datasets and are not publicly accessible. However, a small number of publicly available datasets were identified. The publicly available datasets are highlighted in Table 2.

Table 2. publicly available datasets

Dataset	Samples	Features
Alpha Thalassemia [11]	288	11
PTPP [6]	5066	11
Bangladesh Thalassemia [12]	617	24
Thalassemia [13]	1073	12

3. Proposed System Design

In this section, the study explains the architecture of the proposed system. Figure 1 shows the AI-based Thalassemia diagnosis system. Prior to diagnosis system, To safeguard the integrity and confidentiality of the patient data file, an encryption mechanism was applied before initiating Data Acquisition stage. All files were encrypted using a file-level AES-256-GCM (Advanced Encryption Standard – Galois/Counter Mode) [14] to ensure compliance with data security protocols and to prevent unauthorized access during the research workflow. In the first step of the research workflow, the electronic health record (HER) is retrieved from the medical laboratory clinic. Then, the structure of the electronic blood test report within the electronic health record (EHR) is extracted. The extracted data requires a preparation phase to convert the data (electronic blood test report) to a form that is acceptable as an input to the preprocessing phase. Generally speaking, the electronic blood test report contains information about the patient, such as name, ID number, date of birth, sex, Address, Email, phone number, and so on. In addition, it encompasses report details such as test name, date of sample collection, time of sample collection, date of report, and the name of the laboratory [15]. Moreover, it will include a table that have the test result, which usually shown as a table with variant information columns includes analyte name likewise Hemoglobin, Platelets and so on, Result value measured in number, units of the measured number, reference range which gives the lowest and the highest values that are usually seen in a healthy population, and flag or indicator to mark results as low, normal or High. The first step in the preparation phase is to anonymize the personal information of the electronic record, which means removing any personal information such as name, address, Email, and phone numbers by converting these to a case ID and associating this number with the individual patient's EHR. After the anonymize process, the file is converted to form that is aligned with machine learning standards file; In other words, convert the anonymous file to a CSV type file. Now, the file is ready to feed to the AI detection system, and the second phase starts here. The function of the second phase is to preprocess the data, and this step is very important in machine learning and in model training, testing, as well as evaluation [16]. In the preprocessing phase, certain operations are taken, such as removing duplicate records if they exist, removing errors in values, and handling missing values. Moreover, we filter the data by removing unusual columns features, and noisy fields. Furthermore, the text categories values were converted to numbers, and the numeric features were normalized using standard scaling.

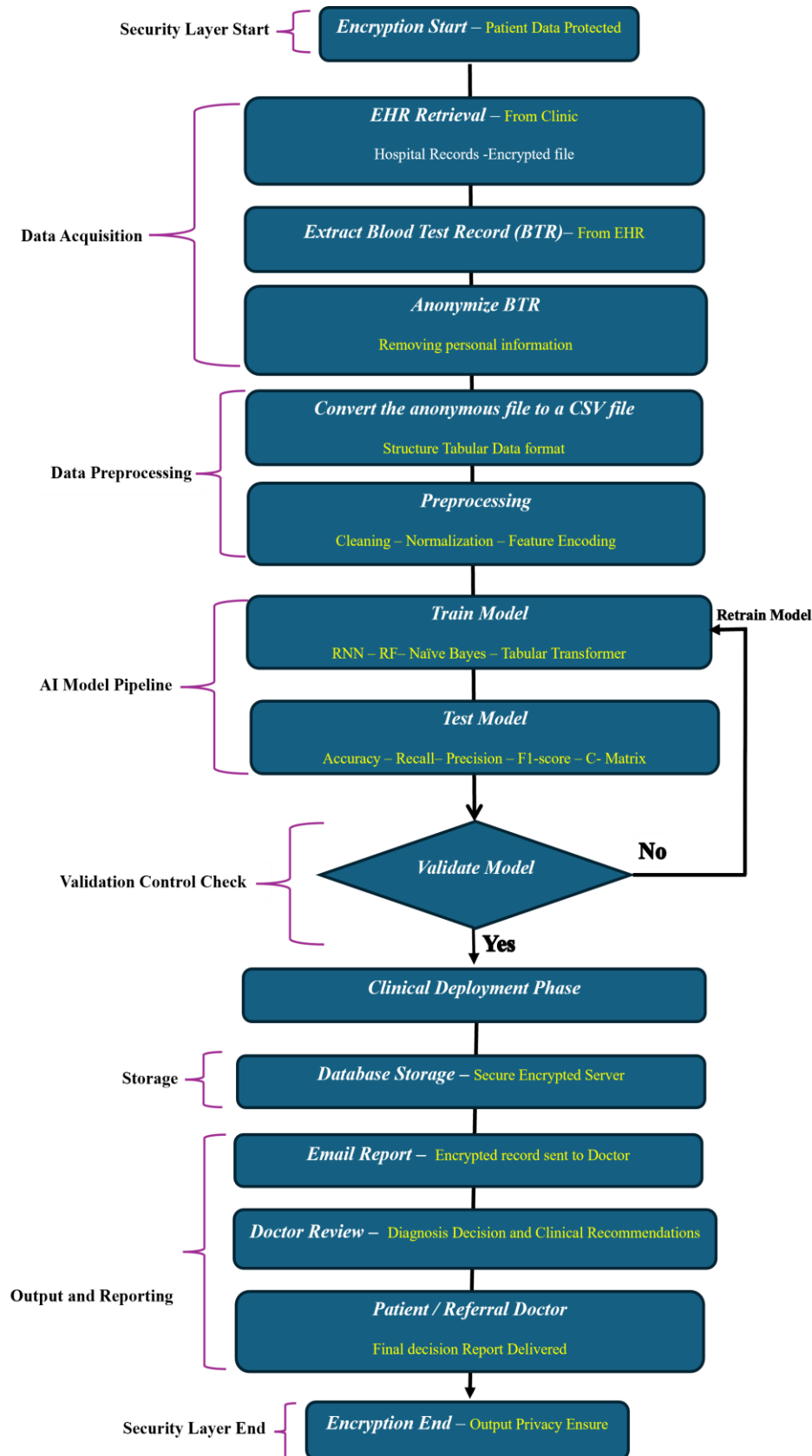


Figure 1: Proposed System Design Diagram

After Preprocessing phase, the preprocessed file is prepared for input to the training phase of the proposed AI detection model. Next, is the training phase in which the system will learn from the prepared processed file the patterns that distinguish healthy and infected patients. Using these labeled and preprocessed samples (prepared from the previous step), a

suitable machine learning model is trained, and its parameters are optimized to maximize detection performance on validation data. The resulting trained model constitutes the decision function that will later be deployed in the inference phase to automatically assess the authenticity of Thalassemia detection decision. The final phase is the test phase; the function of the test set is to take a separate dataset (test) that was not used in training and run the trained model on this unseen data. Then true labels are compared with the model's predictions. Furthermore, calculate the chosen metrics. The final step of the training phase is to decide if the model is good enough and safe to deploy in the real detection environment. When it has been successfully validated, the optimized AI model proceeds to the Clinical Deployment Phase, where it serves as the real-time diagnostic inference engine and does the classification of incoming patient BTR automatically, creating encrypted diagnostic reports for clinicians for final review; hence the use of an AI decision support system in the healthcare process.

Classifier Models and Decision Equations

This section presents an overview of the classification algorithms employed in this study. Each classifier is briefly described, and the decision equation governing its prediction mechanism is defined to provide a formal basis for the experimental evaluation conducted in subsequent sections. The set of classifier models was chosen based on the systematic review conducted, as explained in the introduction section and in association with research question number 4 (RQ4). These classifiers include Naïve Bayes, Random Forest, logistic regression, RNN, and Tabular Transformer. The next section provides a glance at each classifier

3.1 Naive Bayes

Naive Bayes is a simple probabilistic generative classifier based on Bayes' theorem with a conditional independence assumption between features. The predicted class is obtained by maximizing the class posterior as illustrated in equation (1) [17][18]

$$\hat{c} = \operatorname{argmax}_{c \in C} \left(\log P(c) + \sum_{i=1}^n \log P(x_i | c) \right) \dots\dots\dots 1$$

Where each item in the previous equation is defined as follows :

$\hat{c}$: the predicted class label for input x

$x = (x_1, x_2, \dots, x_n)$

x_i : the i -th feature of the input instance

$c \in C$: a candidate class

$P(c)$: prior probability of class c

$P(x_i | c)$: class-conditional probability of feature x_i
given class c

3.2 Random Forest

Random Forest is an ensemble, tree- based discriminative classifier that aggregates the predictions of multiple randomized decision tree via majority voting. The predicted class is obtained by voting over the trees as presented in (2) [19].

$$\hat{c} = \operatorname{argmax}_{c \in C} \sum_{k=1}^K \mathbf{1}(c_k(x) = c) \dots\dots\dots 2$$

k : number of trees in the forest.

$c_k(x)$: class predicted by tree k for input x

$\mathbf{1}(\cdot)$: indicator function, equals 1 if the condition
is true and 0 otherwise

$$\sum_{k=1}^K \mathbf{1}(c_k(x) = c)$$

$\hat{c}$: class with the highest vote count
(majority class)

3.3 Tab Transformer

Tabular Transformer or TabTransformer is a deep neural, Transformer-based discriminative classifier that applies self-attention to embedded tabular features. The predicted class is obtained as in (4) [20]

$$z = f_{\theta}(x)$$

$$P(y = c | x) = \frac{\exp(z_c)}{\sum_{c' \in C} \exp(z_{c'})} \dots\dots\dots 4$$

x : input instance

$f_{\theta}(\cdot)$: neural network with parameters θ

$z \in \mathbb{R}^{|C|}$: logit vector, one score per class

z_c : logit (score) corresponding to class c

$P(y = c | x)$: predicted probability of class c given x

$\hat{c}$: class with the highest predicted probability

3.4 RNN

Recurrent Neural Networks (RNNs) are deep neural sequence models used as discriminative classifiers by encoding sequential inputs into a hidden representation, which is then mapped to class probabilities [21] [22] [23]. Given an input sequence $(x_1, \dots, x_T)$, the RNN produces hidden states $(h_1, \dots, h_T)$. The last hidden state as a sequence representation

$$h_T = \text{RNN}_{\theta}(x_1, \dots, x_T)$$

Then map h_T to class logits $z \in \mathbb{R}^{|C|}$. where

$$z = Wh_T + b$$

After that, convert logits to class probabilities

$$P(y = c | x_{1:T}) = \frac{\exp(z_c)}{\sum_{c' \in C} \exp(z_{c'})} \dots\dots\dots 5$$

Finally, the predicted class with the highest posterior probability as follows

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(y = c | x_{1:T})$$

4. Performance Metrics

This section reports the performance metrics used to evaluate the diagnosis system on the Thalassemia data. The confusion matrix is a visual tool that provides a comprehensive performance evaluation for classification models, revealing how and where a model succeeds or fails. The Confusion matrix measures the overall correctness of the classifier model, diagnosis error types, and serves as the source of all key metrics. In general, there are only four possible prediction outcomes for any sample X from a given class of instances, reflected in the confusion matrix. These encompass the following [24][25]:

1. True positive (TP): Prediction is positive, and the X sample is Thalassemia.
2. True negative (TN): Prediction is negative, and X is healthy.
3. False positive (FP): Prediction is positive, and X is healthy (These are considered Type I errors).
4. False negative (FN): Prediction is negative, and X is Thalassemia (These are considered Type II errors).

According to these outcomes, the model classification quality will be estimated with the following:

$$\text{Precision} = \frac{TP}{TP + FP} \dots\dots\dots 6$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \dots\dots\dots 7$$

$$\text{Specificity} = \frac{TN}{TN + FP} \dots\dots\dots 8$$

$$\text{Recall} = \frac{TP}{TP + FN} \dots\dots\dots 9$$

Finally, the F1-score offers a unified metric that balances precision and recall, providing a comprehensive overview of the model's proficiency in distinguishing between Healthy and Thalassemia instances.

$$\text{F1_score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \dots\dots\dots 10$$

These metrics provide an exhaustive assessment of the classifier's performance. When considered collectively.

5. Results and Discussion

This section presents the experimental results obtained from the proposed Thalassemia-based AI Diagnoses Framework. Table 3 and Figures 2 through 5 summarize the performance of the evaluated classifiers on PPTP dataset. As presented in figure 6, The RNN classifier achieved an accuracy of 96.84%, with a precision of 95.841%, recall of 97.77%, and an F1-score of 96.80%, as derived from the confusion matrix (TP = 484, TN = 498, FP = 21, FN = 11). The low FN count (31) indicates a strong detection rate, while the minimal FP count (21) reflects a satisfactory false disease diagnosis rate.

As illustrated in the confusion matrix Figure 4, the Tabular Transformer classifier achieved a high TP count of 425, indicating strong diagnosis capability of Thalassemia instances of blood test recored instances. The FN value count of (70) observed in the confusion matrix suggests that the model effectively minimizes missed diagnosis, which is of critical importance in Thalassemia based AI Diagnosis applications. Although a highest number of FP of (77) instances were recorded, the overall precision of Tabular Transformer classifier remained high, confirming the model's ability to distinguish Healthy from Diagnosis population. The diagonal elements of the confusion matrix, corresponding to TP count of (425) and TN count of (442), demonstrate that the classifier achieved satisfactory performance across both positive and negative classes.

The confusion matrix of Figure 2 indicates that Naive Bayes exhibits a lower FN count of (25) compared to Tabular Transformer, suggesting superior sensitivity in detecting Thalassemia instances.

The performance of the Random Forest classifier was assessed by means of the confusion matrix based on the data classified by means of the test set, as seen in Figure 3. The matrix shows the Random Forest has received 495 instances of (True Positives) and 516 instances of (True Negatives), resulting in an overall classification accuracy of 99.5%. The diagonal values therefore indicate that the Random Forest classifier is capable of robust classification in both classes.

Regarding the misclassifications, the Random Forest classifier has yielded only 3 False Positive instances and 0 False Negative instances. This means that the Random Forest classifier was able to eliminate the misclassification of Thalassemia patients. The value of False Negatives is of special value because it points towards the patients with Thalassemia who may not get diagnosed in real-life cases.

This study indicates that the Random Forest classifier is able to outperform the existing classifiers and become the most suitable candidate for the project presented. The Random Forest has provided the best classification accuracy of 99.99% with precision of 99.79% and recall of 99.89%, as seen in Table 3. The Random Forest Classifier has managed to outperform the other classifiers with its precision of 99.5% and recall of 99.8% values. These findings suggest that Random Forest is considered an excellent candidate in the Thalassemia-based AI Diagnosis Framework.

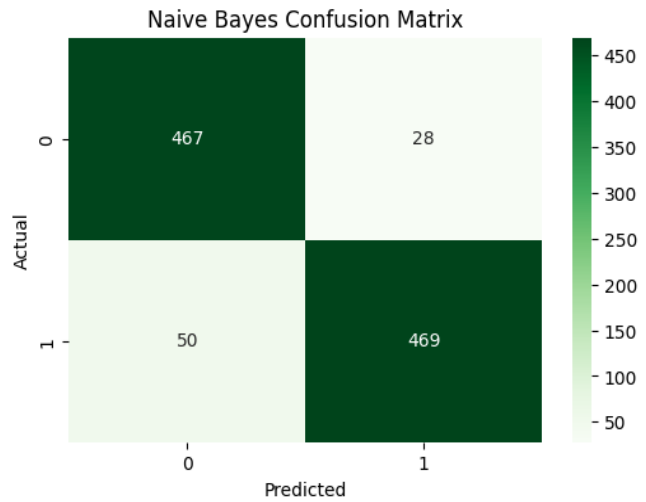


Figure 2 :Confusion Matrix for the Naive Bayes Classifier.

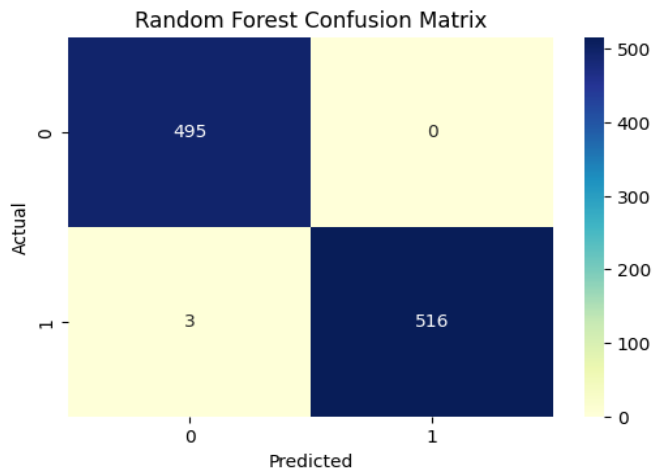


Figure 3 :Confusion Matrix for the Random Classifier.

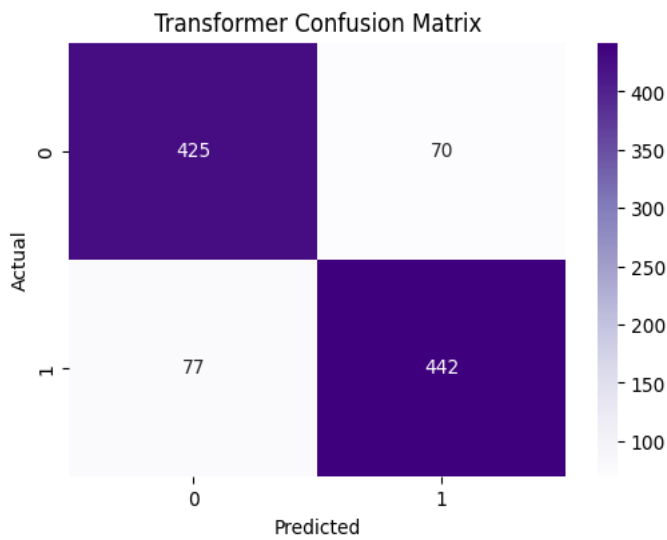


Figure 4 :Confusion Matrix the TabTransformer Classifier.

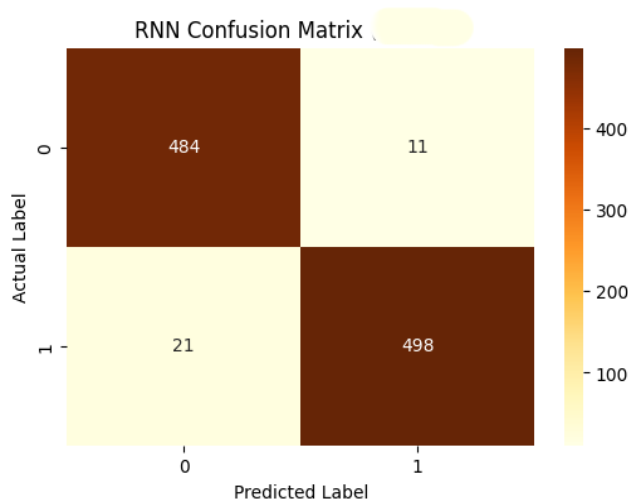


Figure 5 :Confusion Matrix for the RNN Classifier.

Table 3. Performance Evaluation of proposed framework

Classifier	Accuracy	precision	recall	f1-score
RNN	0.972	0.958	0.978	0.968
Random Forest	0.997	0.994	1.000	0.997
Naïve Bayes	0.912	0.903	0.943	0.923
Tabular Transformer	0.835	0.847	0.859	0.853

6. Conclusion

The AI-based Thalassemia diagnosis system was developed in this study to help in enhancing both effectiveness and efficiency in diagnosing Thalassemia. It combines machine learning and deep learning technology with automated preprocessing and feature extraction in order to deal with shortcomings of traditional methods of diagnosis plagued with such problems as human mistakes, the time-consuming nature of the process, and differences in the level of doctors' qualifications. As for the results obtained, they say that the suggested model achieved 100% and 97% effectiveness in the case of the usage of the Random Forest and RNN classifiers, respectively. Thus, it confirms that the developed approach outperforms many traditional methods and is capable of differentiating infected samples from those that are not infected. The results obtained by the system from the PPTP dataset show that the suggested method of diagnosis is very robust and reliable, thus making it possible to apply it in clinics in real-life practice. Finally, this study is important for the field of medical AI since it allows for the provision of a new automated Thalassemia diagnosis system that makes the process of diagnosis much easier and quicker.

References

- [1] Centers for Disease Control and Prevention, "About thalassemia," U.S. Department of Health & Human Services, May 2024. <https://www.cdc.gov/thalassemia/about/index.html> (accessed May 28, 2026).
- [2] Angastiniotis, M., & Lobitz, S. (2019). Thalassemias: An Overview. *International journal of neonatal screening*, 5(1), 16. <https://doi.org/10.3390/ijns5010016>.
- [3] E. Foord, D. Sairafi, M. Berg, S. Westerling, T. Adigun, M. Uzunel, T. Papasavva, and M. Uhlin, "Decoding thalassemia and sickle cell disease: Advances in molecular technologies for comprehensive variant detection," *Front. Genet.*, vol. 17, Art. no. 1810737, 2026.
- [4] Thalassaemia International Federation, "Global thalassaemia review," 2026. <https://thalassaemia.org.cy/what-we-do/global-thalassaemia-review/> (accessed May 28, 2026).
- [5] Centers for Disease Control and Prevention, "Thalassemia," U.S. Department of Health & Human Services, 2024. <https://www.cdc.gov/thalassemia/index.html> (accessed May 28, 2026).
- [6] S. Sadiq *et al.*, "Classification of β -thalassemia carriers from red blood cell indices using ensemble classifier," *IEEE Access*, vol. 9, pp. 45528–45538, 2021.

- [7] M. U. Nasir, M. Zubair, M. T. Naseem, T. Shahzad, A. Saeed, K. M. Adnan, and A. H. Gandomi, "Multiclass classification of thalassemia types using complete blood count and HPLC data with machine learning," *Sci. Rep.*, vol. 15, no. 1, Art. no. 26379, 2025, doi: 10.1038/s41598-025-06594-6.
- [8] F. Rustam, I. Ashraf, and S. Jabbar *et al.*, "Prediction of β -thalassemia carriers using complete blood count features," *Sci. Rep.*, vol. 12, no. 1, Art. no. 19999, 2022, doi: 10.1038/s41598-022-22011-8.
- [9] M. U. Nasir, M. T. Naseem, T. M. Ghazal, M. Zubair, O. Ali, S. Abbas, M. Ahmad, and K. M. Adnan, "A comprehensive case study of deep learning on the detection of alpha thalassemia and beta thalassemia using public and private datasets," *Sci. Rep.*, vol. 15, no. 1, Art. no. 13359, 2025, doi: 10.1038/s41598-025-97353-0.
- [10] H. A. Younas, B. Shoaib Khan, A. H. Khan, A. Bilal, A. Algarni, R. Sarwar, and S. J. Mousavirad, "Prediction of β thalassemia carrier using federated learning and explainable AI," *Front. Med.*, vol. 13, Art. no. 1687773, 2026.
- [11] F. Christensen, D. K. Kılıç, I. E. Nielsen, T. C. El-Galaly, A. Glenthøj, J. Helby, H. Frederiksen, S. Møller, and A. D. Fuglkjær, "Classification of α -thalassemia data using machine learning models," *Comput. Methods Programs Biomed.*, vol. 260, Art. no. 108581, 2025.
- [12] S. S. Ayon, M. E. Hossain, M. S. U. Miah, B. M. Arafin, A. Chowdhury, and N. N. I. Prova, "Thalassemia dataset covering clinical, socioeconomic, and mental health aspects," *Data Brief*, Art. no. 112082, 2025.
- [13] S. S. Abdul-Jabbar and A. Farhan, "Thalassemia dataset," Mendeley Data, 2022, doi: 10.17632/p8rv84hrbs.1.
- [14] National Institute of Standards and Technology, *Advanced Encryption Standard (AES)*, Federal Information Processing Standards Publication (FIPS) 197-upd1, U.S. Dept. Commerce, Washington, DC, USA, 2023, doi: 10.6028/NIST.FIPS.197-upd1.
- [15] *Health Informatics—Requirements for an Electronic Health Record Architecture*, ISO Standard 18308:2011, International Organization for Standardization, Geneva, Switzerland, 2011.
- [16] M. Maharana, S. Mondal, and R. Nemade, "Analyzing the effect of data preprocessing techniques on machine learning-based prediction model outcome," *Health Sci. Rep.*, vol. 6, no. 2, Art. no. e1103, 2023, doi: 10.1002/hsr2.1103.
- [17] T. M. Mitchell, "Naive Bayes and logistic regression," in *Machine Learning*, 2nd ed. (draft chapter), Carnegie Mellon Univ., Pittsburgh, PA, USA.
- [18] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.
- [19] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [20] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "TabTransformer: Tabular data modeling using contextual embeddings," arXiv preprint arXiv:2012.06678, 2020, doi: 10.48550/arXiv.2012.06678.
- [21] Chishti, S., Nosheen, F., Fatima, J., Sultan, N., & Khalid, M. (2025). From CBC to clarity: Interpretable detection of beta-thalassemia carriers in imbalanced datasets. *PloS one*, 20(9), e0331985. <https://doi.org/10.1371/journal.pone.0331985>.
- [22] Subramanian, S., et al. (2025). Detection of β -Thalassemia trait from a heterogeneous population with red cell indices and parameters. *Computers in Biology and Medicine*, 192(Part A), 110151. DOI: 10.1016/j.compbimed.2025.110151
- [23] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data—Recommendations for the use of performance metrics," in *Proc. 2013 Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, 2013, pp. 245–251, doi: 10.1109/ACII.2013.47.
- [24] Ghanim, S. K. (2026). Hybrid ResNet50 and DenseNet121 with adaptive weighted feature fusion in infectious chest diseases classification. *International Journal of Computational and Electronic Aspects in Engineering*, 7(2), 211–221.
- [25] Abdkhaleq, M. H. G. (2025). Hybrid deep learning and neuro-fuzzy approach for COVID-19 diagnosis using CT scan imaging (CNN + ANFIS + PCA). *International Journal of Computational and Electronics Aspects in Engineering*, 6(2), 81–88.