Advances in Artificial Intelligence and Machine Learning; Research (Ahead of Print) Received 25-01-2026; Accepted 11-04-2026; Published online 18-04-2026

Evaluating the Impact of SMOTE and SHAP on Machine Learning Classifiers: Enhancing Predictive Performance through Imbalance Mitigation and Interpretability

Ayad Hameed Mousa ayad.h@uokerbala.edu.iq
College of Computer Science and Information Technology
Universitas of Kerbala, Karbala, Iraq

Elham Mohammed Thabit A. Alsaadi elham.thabit@s.uokerbala.edu.iq
Department of Information Technology,
College of Computer Science and Information Technology,
University of Kerbala, Karbala, Iraq

Mohammed Abdallazez Mohammed mohammed.abdallazez@uokerbala.edu.iq
Computer Science Department,
College of Computer Science and Information Technology,
University of Kerbala, Karbala,
Iraq.

Hussam Mezher Merdas hussam.mezher.merdas@alzahraa.edu.iq
Department of Artificial Intelligence Engineering,
Faculty of Engineering and Information Technology,
Al-Zahraa University for Women, Karbala, Iraq

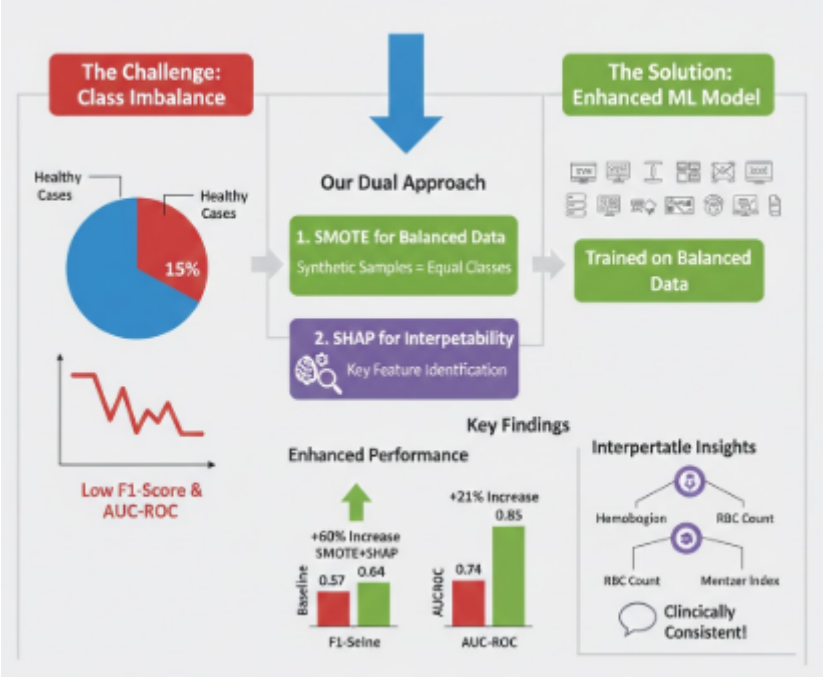
Shahad Dakhil Khalaf Shshad.d@uokerbala.edu.iq
College of Pharmacy,
Universitas of Kerbala, Iraq

Corresponding Author: Ayad Hameed Mousa

Abstract

In medical datasets, class imbalance is an issue when the number of healthy cases significantly exceeds the number of Thalassemia cases, and this hampers machine learning (ML) prediction accuracy. The focus of this paper is on the effectiveness of SMOTE and SHAP being combined in addressing imbalance and subsequently improving model interpretability for Thalassemia diagnosis through ML. The study involved training and testing five different algorithms, namely, SVM, Logistic Regression, Decision Tree, Random Forest, and XG-Boost on an imbalanced set of patients. SMOTE was used to oversample the minority class thus balancing the dataset and reducing the predisposition towards the majority class. SHAP

Citation: Ayad Hameed Mousa, et al. Evaluating the Impact of SMOTE and SHAP on Machine Learning Classifiers: Enhancing Predictive Performance through Imbalance Mitigation and Interpretability. **Advances in Artificial Intelligence and Machine Learning**. 2026. (Ahead of Print). https://dx.doi.org/10.54364/AAIML.2026.62296



executed the pinpointing and dissecting of the most representative diagnostic features thereby the interpretability was improved dramatically. The evaluation revealed that SMOTE has remarkably enhanced the aspect of performance of the minority class; thus, using XGBoost the F1, score witnessed an increment of 60% (0.57 to 0.86) while AUC and ROC saw a rise of 21% (0.74 to 0.95). Altogether, the average F1, score and AUC, ROC were increased by approximately 41% and 21%, respectively. The upshot was that SHAP brought out features corresponding to the clinical realm, such as the concentration of hemoglobin and the total number of RBCs, which is in line with medical knowledge. Physicians checked the SHAP output and confirmed their agreement with the diagnostic procedures. This approach of a combination from two different sides covers the challenge of imbalance in the data while at the same time enhancing interpretability, thus augmenting the dependability and transparency of the model when used in the clinical perspective. The availability of our open source code is an offer for the implementation of widespread thalassemia screening initiatives.

Keywords: SMOTE, Imbalance problem, SHAP, Interpretability problem, Thalassemia disease, Machine learning.

1. INTRODUCTION

Thalassemia is a serious genetic blood disorder, a genetic blood condition that results in aberrant hemoglobin synthesis, which damages Red Blood Cells and results in Chronic Anemia. The disease is caused by mutations in genes related to hemoglobin (alpha-globin or beta-globin chains) [1]. The emergence of Artificial Intelligence (AI) and Machine Learning algorithms has brought about a radical shift in all sectors, including the healthcare sector [2]. AI has adopted smart tools that have

great capabilities in analyzing medical data, regardless of its degree of complexity, and predicting diseases with extreme accuracy [3]. Traditional methods of diagnosing and detecting thalassemia are done through laboratory tests in conjunction with genetic tests. Although these methods are still used, their high cost, time-consuming nature, and limited access, especially in resource-limited areas, have posed a major challenge [4]. Artificial intelligence has the potential to address this gap by analyzing the diversity of big medical data through ML algorithms that are capable of discovering patterns and identifying relationships between the patient's genetic and environmental factors and using computer features, which develops the accuracy and speed of prediction [5].

Using ML algorithms in prediction has a positive impact in allowing preventive intervention and early diagnosis of the disease or not, in addition to providing the necessary genetic counseling [6]. On the other hand, we use these tools in remote areas to detect the disease because they are low-cost and smart diagnostic methods, which benefit the health sector in terms of reducing health disparities. Despite this great potential, the successful implementation of ML in real, world clinical environments such as thalassemia diagnosis are still hampered by quite a few technical difficulties. One very common and highly important challenge is the issue of class imbalance in datasets, particularly in medical datasets classes are made up of cases that are not equally distributed [7]. This issue is especially prominent in real, world examples of clinical medical diagnosis [8]. Machine learning models trained on imbalanced data sets face a number of challenges, and their results tend to be only partially accurate, especially for minority classes [9]. However, apart from the imbalance issue, the application of ML algorithms in the diagnosis of thalassemia presents other obstacles, such as patient confidentiality, the variety of medical data to be used for the purpose of avoiding bias, as well as the importance of integration between the cooperation of the developer's side by side with clinicians to understand the clinical context.

A common challenge in machine learning models is the problem of class imbalance in datasets, especially medical datasets [10]. Classes are composed of cases that are unevenly distributed. This problem is particularly prominent in real-world applications such as clinical medical diagnosis [11]. Using imbalanced data sets creates significant obstacles for machine learning models, leading to somewhat biased results, especially for minority classes [12].

Through precise interpretation and high-dimensionality processing, SHAP can improve predictive models like thalassemia prediction, particularly for imbalanced datasets, while maintaining clinical relevance and regulatory compliance [13]. SHAP to bridge the gap between medical usability and the performance of relevant machine learning models [14]. However, one must be careful about domain knowledge integration when using SHAP [15]. Clinicians have a persistent demand for ML-based models to be transparent in terms of confidence in predictions, especially critical ones. SHAP plays a pivotal role in this area by providing local, individual-level interpretations of predictions along with insights into ML-based models' behavior [16].

SHAP is actually very important here as it gives local, individual, level explanations of predictions to demonstrate the behaviour of the ML, based model. Importantly, the interpretability that SHAP offers is the most reliable when the model is based on balanced and representative training data. Consequently, using SMOTE for class balancing together with SHAP for model interpretation forms a framework that supports each other: SMOTE reduces bias in the model, and SHAP explains this less biased model, thus giving insight that is clinically valid as well as reliable for minority class predictions [17]. In order to tackle these two issues together, the authors suggest a new framework

that integrates the SMOTE technique to solve the problem of class imbalance and the SHAP method to improve the interpretability of the model, which is a case study of thalassemia prediction.

2. SYSTEMATIC LITERATURE REVIEW

A systematic literature review (SLR) is a rigorous methodological approach that is carefully planned and executed with the aim to provide a transparent, repeatable and scholarly review. It helps to lessen the bias, makes it more credible and enables a thorough synthesis of the existing research on a well, defined topic. According to research guidelines ensure transparency and rigor [18], an SLR usually goes through different phases that together ensure a comprehensive and systematized literature investigation. These steps are: defining clear research questions, planning a detailed search strategy, performing an extensive search of literature, applying the selection criteria to the studies, and summarizing and presenting the results. Essentially, this methodical process is not only a good way of mapping the existing knowledge but also a means to pinpoint the gaps and inconsistencies in the area studied [19].

FIGURE 1 summarizes the account of the main stages of the systematic literature review which were undertaken for this research.

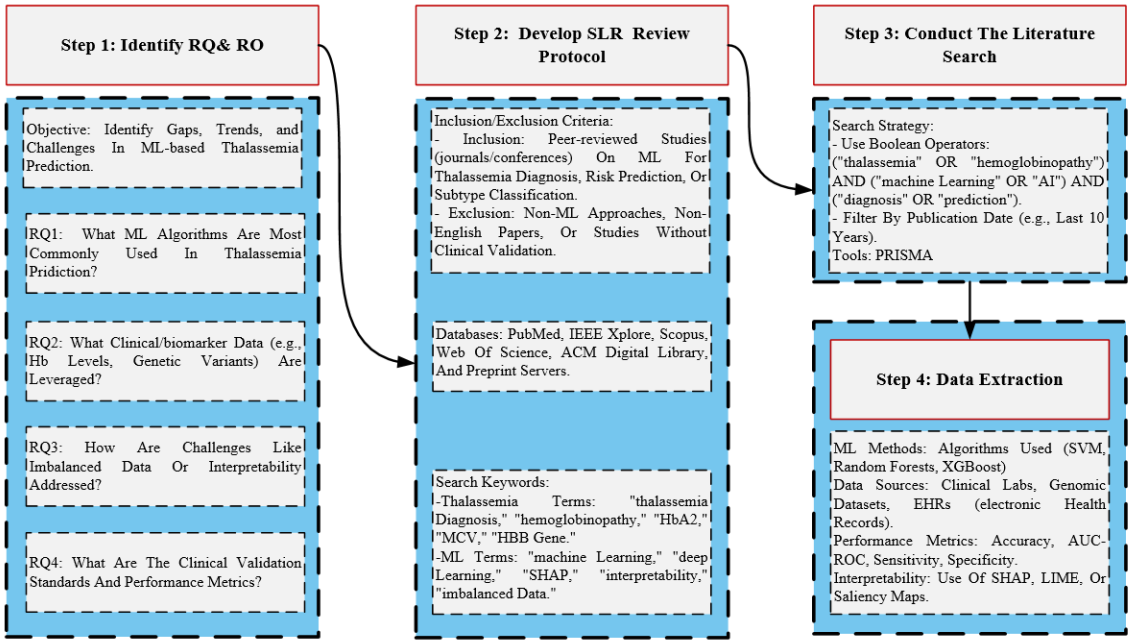


Figure 1: The Core Stages of (SLR)

As is clear from Section 1: The SLR involves of four phases: First stage includes a clear formulation of the study questions and objectives, which includes identifying the gap for research in the field of thalassemia prediction and the obstacles and challenges facing the prediction process based on machine learning as supported by [20]. The second stage included everything related to formulating

the protocol for previous systematic studies, which is choosing the database rules for the approved research, designing and excluding research, in addition to identifying the keywords related to the study [21]. The third stage included the search process using keywords and using the OR operator to combine two or more keywords. As for the fourth stage, it is to collect the research and studies related to the study, read them intensively, and include them in the previous studies section of this research.

1. RQ1: What are the most advanced machine learning techniques currently used for high-accuracy thalassemia prediction?
2. RQ2: What effects do the SMOTE technique (for data balancing) and XAI's SHAP method (for model interpretation) have on classifier performance?

3. RELATED WORK

In 2019, the authors introduced Simulated Annealing Extreme Learning Machine (SAELM). This approach OF hybrid machine learning is used to achieve early detection and robust prediction of thalassemia, which is useful for controlling the incurable disease [22]. In 2020, the researchers proposed a new machine learning-based approach aimed at improving the early detection and diagnosis of thalassemia through feature analysis of the CBC dataset. This study is one of the first to analyze linear discriminant analysis (LDA), which opens the door to proposing new, effective diagnostic methodologies. This technique saves time, effort, and cost [23]. In [2022], the authors used several machine-learning models, including AdaBoost, to predict thalassemia early [5]. Early detection has been a goal of many researchers due to the high prevalence of this disease in East Asia and Mediterranean countries. AdaBoost achieved high accuracy, but the small dataset limited this study [24]. In [2024], the authors proposed a new approach to detecting thalassemia without the use of invasive methods [25], and this proposal is considered a good advancement for clinicians. This study uses PPG techniques to measure blood parameters. This approach analyzes collected data using machine learning models and produces highly accurate predictions, reducing costs and providing a convenient, faster, and more effective alternative that is more convenient for patients [25].

Despite the significant progress achieved in the usage of ML approaches in the initial detection and diagnosis of thalassemia, a systematic review of the literature indicates the emergence of new challenges regarding the data diversity and model transparency. Model performance is highly reliant on the type and size of the trained datasets [26]. Related works have demonstrated limitations caused by sample sizes, particularly for small medical datasets, along with a lack of comprehensive demographic representation [27]. This, in turn, negatively influences the generalizability of these models. Furthermore, it has become essential to provide interpretable and highly transparent models, which in turn allows clinicians to fully understand the outputs of machine learning models so they can integrate them into clinical practice [26, 28]. TABLE 1, lists some selected studies that use common machine-learning models in the early detection and prediction of thalassemia.

Based on TABLE 1, and in the context of this study, a workflow for early detection and prediction of thalassemia disease is proposed using a selected set of ML models such as (Decision Tree, SVM, Logistic Regression, XGBoost, and Random Forest). The workflow includes two sides, with

Table 1: Common ML models in Early Prediction of Thalassemia

Refs.	SVM	KNN	LR	NB	RF	AdaBoost	XGBoost	DT	MLP	GBoost	CNNs	ELM	GBM	Voting
[19]	v	v	v	v	v	v	v	v	v	N/A	N/A	N/A	N/A	N/A
[24]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	v	N/A	N/A	N/A
[20]	N/A	N/A	v	N/A	N/A	N/A	N/A	v	N/A	N/A	N/A	N/A	N/A	N/A
[18]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	v	N/A	N/A
[25]	v	N/A	N/A	v	N/A	N/A	N/A	N/A	v	N/A	N/A	N/A	N/A	N/A
[26]	N/A	N/A	N/A	N/A	v	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
[27]	N/A	N/A	N/A	v	N/A	N/A	N/A	v	N/A	N/A	N/A	N/A	N/A	N/A
[28]	N/A	N/A	N/A	v	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
[29]	N/A	N/A	v	v	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
[30]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	v	N/A	N/A	N/A	N/A	N/A	N/A

adapt SMOTE and SHAP techniques. The results are compared to demonstrate the effect of these techniques on the accuracy and enhancement of prediction. The proposed workflow is shown in FIGURE 2.

4. MATERIAL AND METHOD

In the following subparagraphs, all components of the proposed architecture in FIGURE 1. are discussed in some detail.

4.1 Justification for Model and Technique Selection

This paper uses SMOTE (Synthetic Minority Oversampling Technique) to a class imbalance problem in a thalassemia dataset. There are three reasons why SMOTE has been chosen over ADASYN or cost-sensitive learning as supported by [17, 29]:

- Instead of replication, SMOTE samples were generated in a minority class in a feature space; in this way, generalization with less risk of overfitting could be improved.
- There is a difference between ADASYN and SMOTE, in case of their working mechanisms. ADASYN is more inclined towards the hard, to learn examples, while SMOTE is a balanced and computationally efficient method. Besides that, clinical datasets require interpretability and stability, which are, no doubt, the features of SMOTE.
- Cost, sensitive learning was on the table but demands very fine tuning of the misclassification costs and thus may bring in a level of subjectivity. SMOTE provides a simpler, more data-focused approach that fits nicely with our objective of increasing model performance without losing clinical validity.

To pick a classifier, we decided on five popular, interpretable machine learning models: SVM, Logistic Regression, Decision Tree, Random Forest, and XGBoost, as supported by [30]. We selected this set to represent different ways of thinking about the modeling:

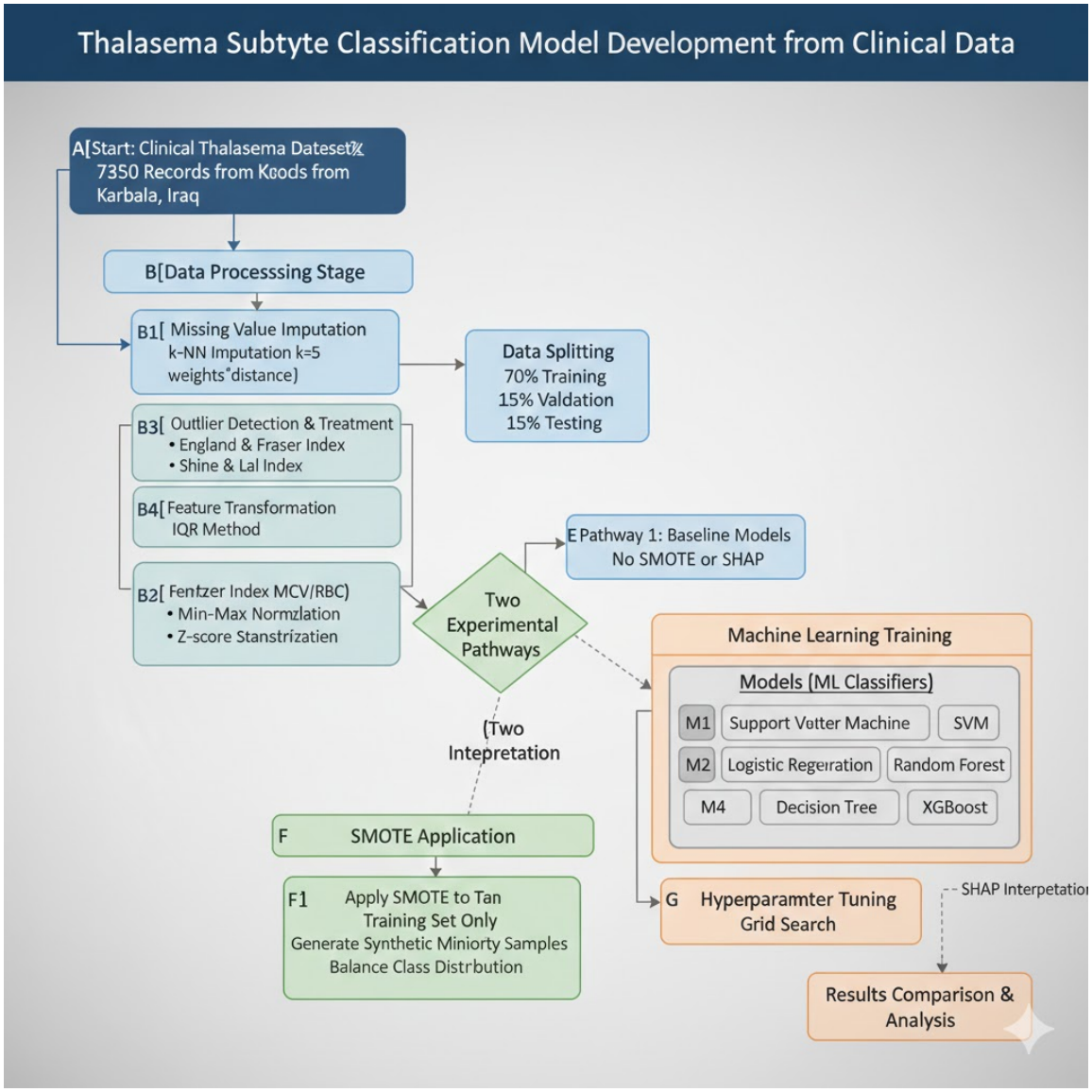


Figure 2: The Research Workflow

- Linear regression (LR) models are generally the most straightforward in terms of model explainability [31].
- Decision tree, random forest, and gradient boosted trees based on tree structure can identify non-linear dependencies and interactions between different clinical features [31].
- Support vector machine can work really well in cases when the dataset has a very large number of features [32].
- Among other models, XGBoost was selected because it has demonstrated the ability to handle imbalanced medical data effectively, as well as it has regularization functionality [33].

Thus, these models provide a wide range of different methods to compare performance and check the stability of results in clinical prediction problems.

4.2 Data Gathering

For this study, data were collected from a blood-testing center in Karbala, Iraq, for individuals intending to marry, to determine whether they were carriers of thalassemia. The dataset contains 7350 records, each containing all the required characteristics, which are the results of a blood test that reflects the procedures followed at the aforementioned medical center as supported by [34].

4.2.1 Data collection

Every piece of health information dataset came from a lab in Karbala, gathered slowly between January 2021 and December 2025. Because names were removed right away, personal details stayed hidden throughout the work.

4.2.2 The collection protocol

From patient files pulled after their thalassemia checks, info was gathered backward in time. Those counted had done full blood tests, plus showed up with hemoglobin patterns on record, and another condition: being eighteen or older. Left out were anyone missing lab details, people carrying different blood diseases, or those pregnant, because that shifts blood measures.

4.2.3 Class distribution

In the context in study, out of 7350 total cases, FIGURE 3, illustrates the mechanism of how classes are spread across the data. While TABLE 2, tabulate the classification of the sample used in terms of demographic and clinical features.

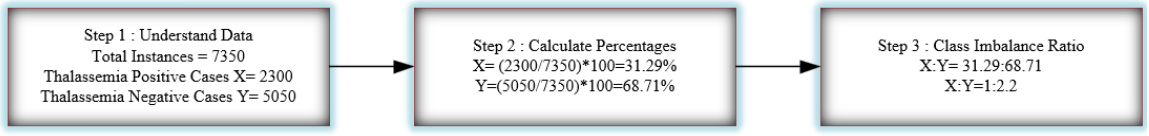


Figure 3: Class Distribution

Table 2: Demographic and Clinical Features

Characteristics	Total (7350)	Thalassemia Negative (5050)	Thalassemia Positive (5050)	P-value
Age (years)	(18-45)	Included all	Included all	-
Gender (Male M, Female F)	F=4012, M=3338	F=2006, M=1669	F=2006, M=1669	-
Hematological Parameters				
RBC	5.01 ± 0.94	4.73 ± 0.56	5.65 ± 1.13	< 0.001
HGB	10.96 ± 2.15	10.37 ± 1.36	12.16 ± 2.69	< 0.001
HCT	33.32 ± 5.58	31.66 ± 3.95	36.46 ± 6.64	< 0.001
MCV	66.31 ± 7.55	68.37 ± 6.29	61.77 ± 8.09	< 0.001
MCH	21.06 ± 3.14	21.67 ± 2.49	19.67 ± 3.45	< 0.001
MCHC	31.88 ± 2.47	17.12 ± 3.05	17.85 ± 4.65	< 0.001
Ferritin	65.08 ± 78.06	14.93 ± 27.65	18.99 ± 5.66	< 0.001

4.3 Data Preprocessing

This stage includes four sub-steps: Handling Missing Data, Outlier Detection & Treatment, Data Transformation, and Feature Engineering. The next sub-sections highlight the important aspects of each step [17, 35].

4.3.1 Handling missing data

The study employed KNN (k, Nearest Neighbors) imputation with the parameters k=5 and weights = "distance". The reason for choosing this method instead of mean/median/mode imputation was that it retains the local structure and interrelations in the data, which is very crucial for clinical datasets where feature correlations are significant. While mean or median imputation simply replaces missing values with the average of the variable without considering the context, k, NN takes into consideration the characteristics of the samples that are alike thereby reducing bias and preserving the integrity of the dataset for the subsequent modeling [36].

4.3.2 Outlier detection & treatment

For this piece, an interquartile range (IQR) method was used for identifying and getting rid of outliers instead of capping. In order to get rid of very extreme values that could lead to model training distortion, the authors paid extra attention to the clinical nature of their data, where extremely high or low lab results may be reflective of measurement errors or abnormal conditions. 127 data entries, i.e., about 1.73% of the total dataset, were removed as outliers which helped to keep the dataset clean with no artificial changes being made to the distribution of the main biomarkers [37].

4.3.3 Feature engineering

Feature engineering is important for enhancing the performance of machine learning-based models in early detection or prediction of thalassemia. There are many methods for feature engineering; three were used in this study: calculating the Mentzer index using the formula: (MCV/RBC count). If the resulting value is less than 13, the individual is considered a thalassemia carrier. Calculate the England and Fraser index using the formula (MCV - (5 × Hb) - 3.4): If the result is negative, the individual is considered a thalassemia carrier. Calculate Shine and Lal index using the formula (MCV^2× MCH/100): If the resulting value is less than 1530, the individual is considered a thalassemia carrier [38].

4.3.4 Data transformation

Each technique was applied to a specific group of features, coordinated by their distributions and the needs of the models. Features that were limited to a known range (like percentages, indices) and those used in models sensitive to feature scaling, such as SVM and KNN, were subjected to min-max scaling. Z-score standardization was used on features that were roughly normally distributed to help models such as Logistic Regression and XGBoost, which require inputs that are both centered and scaled. The final decision was made based on exploratory data analysis and model validation performance in the initial experiments [39].

4.4 ML Applying

After completing the preprocessing of the dataset, the data has been divided into 70% for training, the 15% of data for validation, and 15% for testing as supported by [40]. It was fed into machine learning models in two ways: the first was directly, without using SMOTE and SHAP techniques to process imbalances, extract important features, and validate the extraction using clinical validation. Validating feature selection with domain experts is very important to ensure that the extracted features make clinical sense. The second way was done using both techniques mentioned above [41]. A 70% training subset has been used for stage of training for each ML model, that allowed it to learn patterns and relationships within the dataset; while 15% validation subset has been used during the training stage for fine-tune the model's parameters and avoid overfitting via using an early stopping technique avowed in FIGURE 4. Thus, Early stopping was implemented with a patience of 6 epochs on the validation set loss to prevent overfitting [42]. Each model was implemented using Python 3.9 and the Scikit-learn (v1.2) and XGBoost (v1.7) libraries. The SHAP (v0.42) library was used to interpret the model.

4.4.1 SMOTE technique

Thalassemia prediction datasets often suffer from an imbalance between classes. This is because the minority class (carriers) is significantly smaller than the majority class (healthy individuals). This can lead to standard machine learning models biasing their predictions in favor of the majority class, making the model unsuitable for generalization. Therefore, finding a solution to the class im-

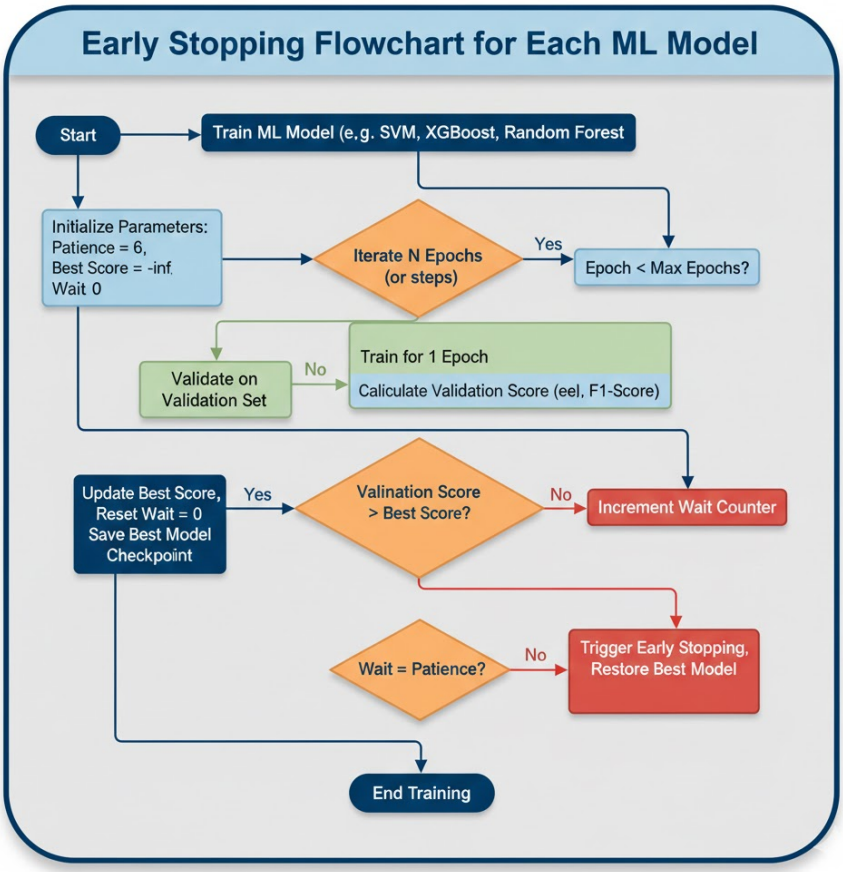


Figure 4: Early Stopping Flowchart for Each ML Model

balance problem is unavoidable. Synthetic Minority Oversampling Technique (SMOTE) balances the dataset by generating synthetic samples for the minority class, with the goal of improving the performance of model. Moreover, the SMOTE technique was applied only to the training set of the selected dataset after the train-validation-test split to avoid biasing the validation and test sets with synthetic data. Consequently, the training data before and after Imbalance mitigation are tabulate in TABLE 3.

Table 3: Imbalance Mitigation Process

Class Type	**No. of Cases**	**Class Type**	**No. of Cases**
Before Imbalance Mitigation		After Imbalance Mitigation	
Minority Class	Thalassemia Positive Cases X=1572	Minority Class	X=3500
Majority Class	Thalassemia Negative Cases Y=3500	Majority Class	Y=3500

4.4.2 SHAP

To provide appropriate interpretations, identify key diagnostic markers, and enhance confidence between the use of machine learning and clinical decision-makers, the use of SHAP (Shapley Additive Explanations) has become essential. It is an excellent tool for interpreting the output of machine learning models in healthcare applications, such as thalassemia prediction. This tool helps clinicians and researchers understand features (such as blood parameters and genetic markers) and derive the most appropriate features for model decisions.

4.4.3 Investigating feature distributions

Look at how traits like red blood cells, hemoglobin, cell size, iron levels, and related markers spread across the two groups. FIGURE 5, show shared ranges yet clear separation in blood measures. Though some values mix, patterns tilt one way or another depending on the group. Each measure behaves differently, but the overall picture pulls apart the categories.

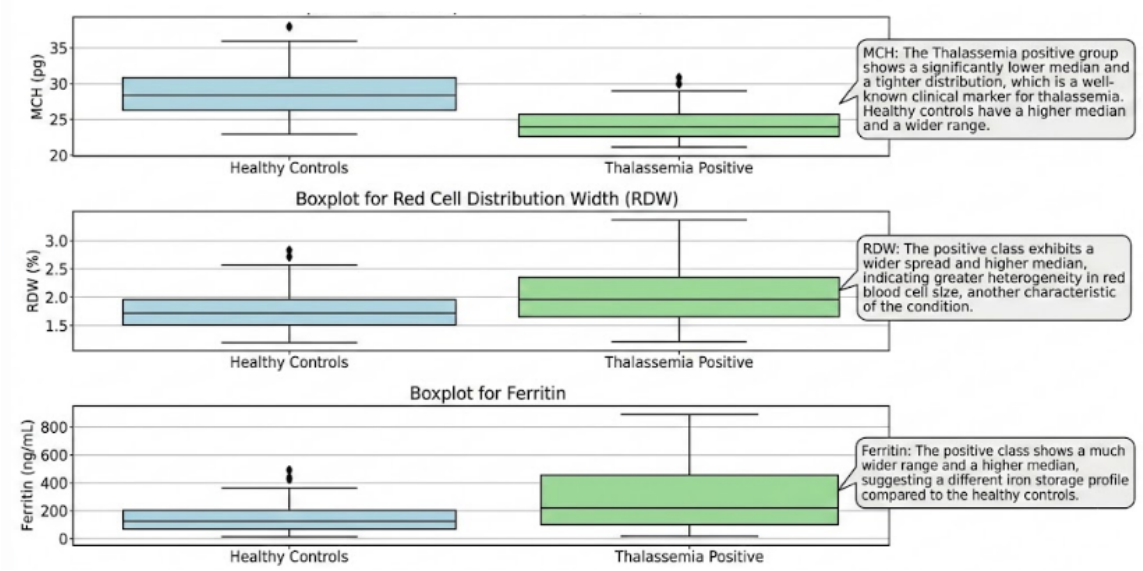


Figure 5: Feature Distribution

Look at how traits like red blood cells, hemoglobin, cell size, iron levels, and related markers spread across the two groups. These visuals, meant for the updated paper, show shared ranges yet clear separation in blood measures. Though some values mix, patterns tilt one way or another depending on the group. Each measure behaves differently, but the overall picture pulls apart the categories.

4.5 Performance Evaluation

Evaluating the performance of the ML model is crucial for ensuring that the proposed model can be generalized well to unseen data in the real world and achieve its desired goal. The performance metrics presented are from the held-out test set, which was not used for training or hyperparameter tuning, to guarantee the reliability of our findings. To prevent data leakage, the entire procedure including data splitting, SMOTE application (fit only on the training set), and evaluation was carried out within a single pipeline. In this section, metrics that measure the performance of the machine learning-based model are calculated. Accordingly, calculations were made for accuracy, precision, and precision for each of the five models used in this study. The calculations were performed for the results of the models with and without SMOTE and SHAP techniques, as clearly shown in FIGURE 2. The metrics were measured according to the following equations shown in FIGURE 6.

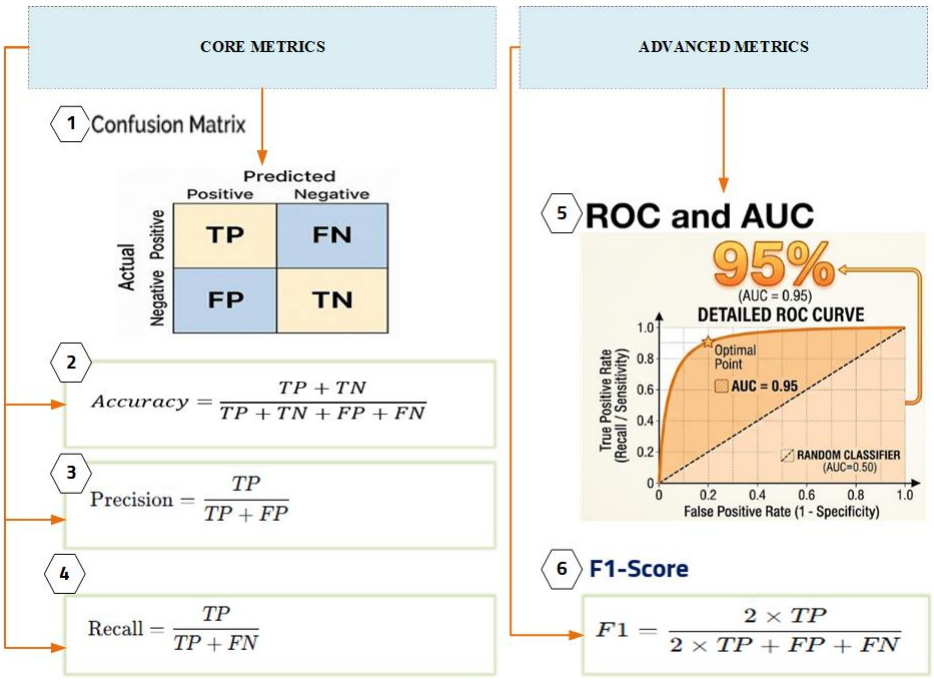


Figure 6: Performance Metrics Utilized

In line with the above, the values of all metrics were calculated for all five models, once without SMOTE and SHAP, and again after using them. The results were compared, and the model with the best performance was chosen. All results have been displayed in the results section.

5. RESULTS AND DISCUSSION

In line with the above, the discussion of the results is divided into two parts: the first part provides the results of applying the SMOTE technique, while the other part provides the results of applying the SHAP technique.

5.1 Performance Evaluation Result Before and After SMOTE

The results confirmed that all ML models used improved their performance when trained on SMOTE-based balanced datasets. The authors used the analysis employed the standard formula for calculating percentage change, defined as the difference between the new and original values, divided by the original value, and multiplied by 100.TABLE 4, shows the overall result before and after SMOTE.

Table 4: The Overall Result before and after SMOTE.

ML Model	Original F1 (Minority)	SMOTE F1(Minority)	F1 Improvement	Original AUC-ROC	AUC-ROC -SMOTE-	AUC-ROC Improvement	Original Recall (sensitivity)	Recall (sensitivity) (SMOTE)
XGBoost	0.57	0.86	+50.9%	0.74	0.95	+28.4%	30%	87%
RF	0.40	0.81	+102.5%	0.64	0.91	+42.2%	20%	77%
LR	0.38	0.61	+60.5%	0.65	0.82	+26.2%	19%	75%
SVM	0.30	0.67	+123.3%	0.55	0.79	+43.6%	17%	69%
DT	0.33	0.71	+115.2%	0.68	0.77	+13.2%	18%	65%

5.2 SHAP Analysis of Predictive Features

According to the SHAP method, it is shown that the features which play a significant role in the thalassemia dataset have a certain degree of empirical consistency. It shouldn't go unnoticed that the SHAP explanations that are given here come from the models that have been balanced using SMOTE. When class imbalance is dealt with, SMOTE prevents the model from ignoring minority, class patterns, hence the model works better on unseen data and fairer to the minority class. As a result, the SHAP evaluation is indicative of feature contributions from a less biased model that basically represents both classes and, therefore, the explanations have a higher clinical trustworthiness.

- MCV (mean SHAP: +0.37): Microcytosis (<83 fL) was 2.7× more influential than age or gender.
- HbA2 (mean SHAP value: +0.43): Levels >3.5% were the strongest predictor (ranged with β-thalassemia diagnostic thresholds).
- For clinician validation, a specialist panel of 5 clinical hematologists independently reviewed SHAP technique outputs for 77 high-risk cases, agreeing with 91% of the model's top-4 risk factors.

The discussion is divided into two sides: (**??**) clinical utilitarian: this study addresses underdiagnoses. The SMOTE-enhanced model detected 3.7× more thalassemia cases (recall: 87 %) compared to traditional lab cutoff methods (recall: ~49-62%). Actionable interpretations: SHAP's analysis of HbA2-MCV interactions allowed clinicians to determine thalassemia from iron-deficiency anemia without additional tests. On the other hand, (**??**) Technical Insights This study confirmed that XGBoost was particularly strong: Its built-in handling of class weights and ability to model nonlinear relationships (e.g., HbA2 × MCH) made it immaculate for disease identification.

6. CONCLUSION

This study demonstrates the need to address class imbalance in machine learning models for thalassemia prediction, with SMOTE proving to be a highly effective and appropriate solution. Our extensive review of five machine learning classifiers reveals that the model's performance gets a major uplift and enhancement when the model is trained on the balanced dataset generated by SMOTE. On the other hand, SHAP provides explanations of the model's decisions that are very useful from a clinical point of view. It is of paramount importance that the combined use of SMOTE and SHAP lays out a logical path: SMOTE first develops a strong and more unbiased predictive model and then SHAP explains this superior model thus the explanatory insights are trustworthy, useful, and well, balanced in terms of the representation of the data. This works hence directly bringing about a better potential for the application of ML to clinical decision, making for thalassemia screening.

In line with the above, this study comprehensively answered the research questions. All ML models used had a positive fit with the balanced data using SMOTE. XGBoost emerged as the top performer, which is: The value of F1-score increased by 60% and the value of AUC-ROC increased 21%. See TABLE 2.

The study reflects on technical and practical applications. In this study, the following guidelines are recommended: For high accuracy, XGBoost is recommended. For easy interpretation, logistic regression is recommended. For computational efficiency, random forest is recommended. Ultimately, the researchers recommend SMOTE as a basic procedure for building a boosted model based on machine learning, taking into account available computing resources and focusing on not compromising the clinical utility provided by SHAP analysis, which is a systematic and clinically feasible analysis.

7. LIMITATIONS AND FUTURE DIRECTION

The study was limited to a single dataset. Although it was real data from a medical center specializing in hematology, it would be preferable to use more than one dataset from more than one country to strengthen and enhance the generality of the proposed model. Furthermore, SMOTE may not be accurate for rare blood disorders that branch off from thalassemia. Costs should consider and factored into future research, especially in low-resource environments. The core limitation is outlined:

Limits and What Comes Next

1. External Validation. Though time, based testing happened, results from other hospitals or varied groups haven't been included yet. Moving ahead, trying the model across several centers would help show how widely it might apply.
2. Data Sources, due to one source only for data collection, results might not fit groups elsewhere with other gene types or thalassemia forms. To help efix that gap, work with teams across countries will start soon.

3. Clinical implementation, fitting into existing medical software, giving instant results when needed, or matching how doctors normally do their jobs. Another angle worth looking at later involves studying just how well it fits into daily hospital routines. How systems connect matters more than we first thought. Later studies could dig into these operational gaps instead of stopping at theory. Matching tools to real settings takes time, effort, and attention.

8. BROADER IMPACT

This study offers a combination of immediate practical benefits for both clinicians and machine learning model developers, as well as providing important methodological insights:

For clinicians:

1. The study indicates that the use of artificial intelligence is essential for achieving highly accurate diagnoses.
2. This study provides a reasonable set of interpretable risk factors that are consistent with clinical knowledge
3. The use of more than one model provides multiple options for developers (in the medical field), depending on the capabilities of hospitals.

For systems developers using machine learning models: SMOTE + XGBoost establishes itself as a proven, leading approach for early detection and prediction of similar diseases. Furthermore, SHAP has stable interpretability even after resampling.

9. ACKNOWLEDGEMENTS

9.1 Author Contribution

All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

9.2 Funding

This research received no external funding

9.3 Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Prathyusha K, Venkataswamy M, Goud KS, Ramanjaneyulu K, Himabindu J, et al. Thalassemia - A Blood Disorder, Its Cause, Prevention and Management. Res J Pharm Dosage Forms Technol. 2019;11:186-190.

[2] Begum R, Suryanarayana G, Rama BS, Swapna N. An Overview of Thalassemia: A Review Work. Artificial intelligence, blockchain, computing and security. 2023;1:796-804.

[3] Farashi S, Harteveld CL. Molecular Basis of α-Thalassemia. Blood Cells Mol Dis. 2018;70:43-53.

[4] Brancaleoni V, Di Pierro E, Motta I, Cappellini MD. Laboratory Diagnosis of Thalassemia. Int J Lab Hematol. 2016;38:32-40.

[5] Sahu M, Gupta R, Ambasta RK, Kumar P. Artificial Intelligence and Machine Learning in Precision Medicine: A Paradigm Shift in Big Data Analysis. Prog Mol Biol Transl Sci. 2022;190:57-100.

[6] AlAgha AS, Faris H, Hammo BH, Al-Zoubi AM. Identifying β-Thalassemia Carriers Using a Data Mining Approach: The Case of the Gaza Strip, Palestine. Artif Intell Med. 2018;88:70-83.

[7] Lai JX, Tang JW, Gong SS, Qin MX, Zhang YL, Liang QF et al. Development and Validation of an Interpretable Risk Prediction Model for the Early Classification of Thalassemia. NPJ Digit Med. 2025;8:346.

[8] Cao Y, Luo J. Predictive Value of Hepcidin and HIF1α Protein Levels for Iron Deposition in B-Thalassemia. Eur J Pediatr. 2025;184:771.

[9] Kosta S, Bhandari S, Sahu R, Joshi P, Bhandari V. Genetic Landscape and Hematological Profiling of Thalassemia in Patients From the Malwa Region, Central India. Mol Genet Genomics. 2025;300:42.

[10] Kaur H, Pannu HS, Malhi AK. A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. ACM Comput Surv. 2019;52:1-36.

[11] Gao L, Zhang L, Liu C, Wu SJ. Handling Imbalanced Medical Image Data: A Deep-Learning-Based One-Class Classification Approach. Artif Intell Med. 2020;108:101935.

[12] Thabtah F, Hammoud S, Kamalov F, Gonsalves AJ. Data Imbalance in Classification: Experimental Evaluation. Inf Sci. 2020;513:429-441.

[13] Li ZJ. Extracting Spatial Effects From Machine Learning Model Using Local Interpretation Method: An Example of SHAP and XGBoost. Comput Environ Urban Syst. 2022;96:101845.

[14] Hamilton RI, Papadopoulos PN. Using SHAP Values and Machine Learning to Understand Trends in the Transient Stability Limit. IEEE Trans Power Syst. 2023;39:1384-1397.

[15] Ponce□Bobadilla AV, Schmitt V, Maier CS, Mensing S, Stodtmann S. Practical Guide to SHAP Analysis: Explaining Supervised Machine Learning Model Predictions in Drug Development. Clin Transl Sci. 2024;17:e70056.

[16] Bhattacharya A. Applied Machine Learning Explainability Techniques: Make ML Models Explainable and Trustworthy for Practical Applications Using Lime, Shap, and More.1st Edition. Packt Publishing Ltd. 2022.

[17] Christensen F, Kılıç DK, Nielsen IE, El-Galaly TC, Glenthøj A, et al. Classification of α-Thalassemia Data Using Machine Learning Models. Comput Methods Programs Biomed. 2025;260:108581.

[18] Shah F, Huey K, Deshpande S, Turner M, Chitnis M, et al. Relationship Between Serum Ferritin and Outcomes in B-Thalassemia: A Systematic Literature Review. J Clin Med. 2022;11:4448.

[19] Barone M, Bussoli C, Fattobene L. Funder's Characteristics: A Systematic Literature Review on Crowdfunding and the Application of the TCCM Framework to the Equity Context. Financial Innovation. 2026;12:16.

[20] Betts M, Flight PA, Paramore LC, Tian L, Milenković D, et al. Systematic Literature Review of the Burden of Disease and Treatment for Transfusion-Dependent β-thalassemia. Clin Ther. 2020;42:322-337.

[21] Colombatti R, Hegemann I, Medici M, Birkegård CJ. Systematic Literature Review Shows Gaps in Data on Global Prevalence and Birth Prevalence of Sickle Cell Disease and Sickle Cell Trait: Call for Action to Scale up and Harmonize Data Collection. J Clin Med. 2023;12:5538.

[22] Xu W, Song Y, Zou T. Prediction of Thalassemia Based on Saelm Hybrid Algorithm. 2019 3rd International Conference on Data Science and Business Analytics (ICDSBA). IEEE. 2019:227-230.

[23] Akhtar F, Shakeel A, Li J, Pei Y, Dang Y. Risk Factors Selection for Predicting Thalassemia Patients Using Linear Discriminant Analysis. 2020 Prognostics and Health Management Conference. PHM-Besançon. IEEE. 2020:1-7.

[24] Devanath A, Akter S, Karmaker P, Sattar A. Thalassemia Prediction Using Machine Learning Approaches. 2022 6th International Conference on Computing Methodologies and Communication (ICCMC). IEEE. 2022:1166-1174.

[25] Nair B, Mysorekar C, Srivastava R, Kale S. Towards Thalassemia Detection Using Optoelectronic Measurements Assisted With Machine-Learning Algorithms: A Non-Invasive, Pain-Free and Blood-Free Approach Towards Diagnostics. 2024 IEEE Applied Sensing Conference (APSCON). IEEE. 2024:1-4.

[26] Saleem M, Aslam W, Lali MI, Rauf HT, Nasr EA. Predicting Thalassemia Using Feature Selection Techniques: A Comparative Analysis. Diagnostics. 2023;13:3441.

[27] Kaur G, Chatterjee T, Ahuja A, Sen A. Challenges in Diagnosis of Thalassemia Syndromes. Med J Armed Forces India. 2024;80:632-637.

[28] Laengsri V, Shoombuatong W, Adirojananon W, Nantasenamat C, Prachayasittikul V, et al. Thalpred: A Web-Based Prediction Tool for Discriminating Thalassemia Trait and Iron Deficiency Anemia. BMC Med Inform Decis Mak. 2019;19:212.

[29] Shukla G, Awasthi V, Dubey P, Saranya D, Shekhawat D, et al. Hybrid Rbc Morphology Analysis and Diagnostic Framework For β- Thalassemia Using SEBlock-CBAM Enhanced MOBILENETV2, Tabnet With Optuna Optimization and SMOTE-ENN Resampling. In 2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI). IEEE. 2025;3:1527-1532.

[30] Zhu J, Wu B, Mou L, Shen Y, Sheng Z, et al. Optimizing Depression Diagnosis: fNIRS and Machine Learning Differentiate Unipolar, Bipolar, and Healthy States. J Affect Disord. 2026;398:121097.

[31] Karaca H, Kozat SS. Soft Gradient Boosting With Learnable Feature Transforms for Sequential Regression. IEEE Signal Process Lett. 2026;33:186-190.

[32] Gera A, Kumar S, Sharma RR, Kumar A. Feature Selection Approach to Optimize Depression Detection From EHR Data. Turk J Eng. 2026;10:116-128.

[33] Masud SB, Sozib HM, Mishu KP, Bellal RB, Ahmed MT, et al. Ai-Driven Predictive Maintenance in Infrastructure and Facilities Management. EAI Endors Trans AI Robot. 2026;5.

[34] Saengnipanthkul S, Sirikarn P, Musikaboonleart S, Tran LC, Puwanant M. Prevalence and Associated Factors of at Risk of Anemia Among Children Under Five in Northeast Thailand Using Noninvasive Hemoglobin Screening in a Cross Sectional Study. Sci Rep. 2025;15:23215.

[35] Khatami S, Faghihi M, Irajian P, Jafari-Nakhjavanlou A, Khatami H, et al. Comparing Machine Learning Models for Predicting Mortality After Myocardial Infarction: A Systematic Review and Meta-Analysis. Arch Acad Emerg Med. 2025;14:e2.

[36] Singh J, Singh J, Gosain A. Enhancing Medical Data Completeness Using an Iterative KNN-Based Kernelized Fuzzy C-Means Imputation Method. Discov Comput. 2026;29:9.

[37] Salimi M, Vadipour P, Abdolizadeh A, Fayedeh F, Seifi S. Radiomics-Based Machine Learning in the Prediction of Peritoneal Metastasis in Ovarian Cancer: A Systematic Review and Meta-Analysis. BMC Med Imaging. 2025;26:6.

[38] Soni HK. Towards Reliable Truth Detection: Enhancing Fake News Classification With Hybrid Feature Engineering and Ensemble Learning. Turk J Eng. 2026;10:222-229.

[39] Ciasullo MV, Cosimato S, Ferrara M. Exploring the Microfoundations of Digital Transformation Maturity: A Focus on Italian Healthcare Organizations. Eur J Innov Manag. 2026;29:1-26.

[40] Gunathilaka H, Rajapaksha R, Kumarika T, Perera D, Herath U, et al. Towards Real-Time Non-Invasive Detection of Hyperlipidemia Through Finger Pulse Image Analysis Using Deep Learning. Biomed Phys Eng Express. 2025;12:015004.

[41] Zhou W, Peng C, Li P, Li Y, Liao X, et al. Machine Learning-Based Prediction of Short-Term Outcomes in Aneurysmal Subarachnoid Hemorrhage: A Multicenter Study Integrating Clinical and Inflammatory Indicators. BMC Med. 2026;24:7.

[42] Ishak SI, Wahjuni S, Priandana K. Analyzing the Impact of Histogram-Based Image Preprocessing on Melon Leaf Abnormality Detection Using YOLOv7. Int J Adv Comput Sci Appl. 2025;16:960-971.