Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Machine Learning-Based Prediction of Thalassemia: A Review

Dawlat Abdulkarim Ali[1,2], Adnan Mohsin Abdulazeez[1]
dawlat.abdulkarim@auas.edu.krd,Adnan.mohsin@dpu.edu.krd
[1] Duhok Polytechnic University
[2] University for Applied Science-Technical College of Informatic-Akre-Department of Information Technology-Kurdistan Region-Iraq

Article Information

Submitted : 16 May 2024
Reviewed : 5 Jun 2024
Accepted : 15 Jun 2024

Keywords

Thalassemia, Machine Learning, Diagnosis, Prediction.

Abstract

This article presents a comprehensive systematic review of recent advancements in machine learning (ML) applications for diagnosing Thalassemia, a genetic hematologic disorder. Focusing on studies from the last five years, this review highlighted significant technological advancements in ML, including the use of predictive modeling, image analysis, and deep learning algorithms, which have considerably improved the accuracy and efficiency of Thalassemia diagnosis. The review evaluates the application of various ML models in analyzing extensive biomedical data, which significantly enhances patient management and treatment outcomes. Key challenges such as data diversity, model transparency, and the need for robust training datasets are discussed, along with the integration of ML into existing clinical workflows. The potential transformative impact of ML in hematology is underscored, critically evaluating its effectiveness and ongoing developments in the field. This review aims to provide insights into the current research trends and future directions in the use of ML for the diagnosis and management of Thalassemia and other similar hematological disorders.

A. Introduction

Thalassemia represented a group of genetic hematologic disorders characterized by anomalies in hemoglobin synthesis, leading to anemia. The disorder is categorized broadly into Alpha Thalassemia and Beta Thalassemia, each attributed to mutations that impair alpha and beta globin chain production, respectively. These conditions manifest in varying degrees of severity, influencing the quality and quantity of hemoglobin and thus the oxygen-carrying capacity of the blood [1], [2]. The complexity of Thalassemia and its clinical implications necessitate advanced, accurate diagnostic strategies. ML technologies offer promising enhancements in the prediction and diagnosis of Thalassemia by analyzing extensive biomedical data, which can significantly improve patient management and treatment outcomes.

The application of machine learning in healthcare, especially in diagnosing genetic disorders like Thalassemia, leverages computational models to interpret complex data sets effectively. This approach includes predictive modeling and image analysis, which are pivotal in recognizing disease patterns and improving diagnostic accuracy [3], [4], For instance, machine learning models have successfully been applied to distinguish between Thalassemia and similar hematological disorders by analyzing blood sample images and genetic data [5], [6].

Recent advancements in deep learning, a subset of machine learning, have introduced sophisticated algorithms capable of diagnosing Thalassemia from medical imaging, such as high-resolution blood smear images. These models can detect subtle morphological changes associated with the disorder, facilitating early and accurate diagnoses, which are crucial for effective treatment. Additionally, the integration of machine learning with existing clinical workflows has demonstrated potential to enhance diagnostic procedures, offering faster and less invasive alternatives to traditional methods [7], [8].

A systematic review of recent advancements in ML applications for diagnosing Thalassemia, focusing on studies from the last five years to include the latest technologies and methodologies, is presented. Key technological advancements are highlighted, and challenges such as data diversity, model transparency, and the need for robust training datasets are addressed. The transformative impact of ML in hematology, particularly in managing Thalassemia, is underscored by critically evaluating its effectiveness and the ongoing developments in the field.

B. Research Method

1 Literature Search and Selection

A comprehensive review focusing on studies published within the last five years was conducted. Key search terms included "thalassemia", "machine learning", "diagnosis", and "prediction."

Peer-reviewed research articles that specifically utilized machine learning for predicting Thalassemia were included. Exclusions were made for non-English articles, conference abstracts, and unrelated studies.

2 Data Extraction and Analysis

Key data such as study objectives, algorithms used, sample sizes, and main findings were systematically extracted and tabulated.

3 Critical Evaluation

Each study was evaluated for methodological soundness and bias. Gaps in current research were identified, leading to suggestions for future studies.

This methodology ensures that the review is comprehensive, up-to-date, and critically engaged with current research trends in machine learning applications for Thalassemia prediction.

C. Litreature Review

Devanath et. al. in 2022 explored the use of machine learning algorithms to predict Thalassemia, a genetic disorder that causes anemia, highlighting the need for early detection due to its prevalence in Asia and the Mediterranean. Their study tests several algorithms, including AdaBoost, which achieved the highest accuracy at 100%. The research aims to improve Thalassemia prediction to enhance treatment strategies, though it acknowledges limitations such as the small dataset size, suggesting future enhancements could include advanced algorithms and a larger dataset [9].

Zaylaa et. al. in 2022 introduced an AI framework that uses Deep Learning for diagnosing Thalassemia through medical imaging. This innovative approach involves a supervised semantic image segmentation model enhanced by data engineering techniques such as annotation, augmentation, and preparation, with a key method being Prediction Time Augmentation (PTA) which improves prediction accuracy and image smoothness. Aiming to surpass the limitations of costly and skill-intensive traditional screening methods like High Performance Liquid Chromatography (HPLC) and DNA testing, the framework achieved a mean Intersection Over Union (IoU) score of 88% with PTA, demonstrating its efficacy. This AI-integrated method promises a more accessible, automated, and cost-effective diagnostic process, potentially transforming Thalassemia detection and healthcare services by enabling quicker and more accurate diagnoses [10].

Binu Nair et al. Introduced a groundbreaking method for diagnosing Thalassemia through non-invasive and pain-free techniques, marking a significant advancement in clinical diagnostics. Their study utilizes photoplethysmography (PPG), a light-based sensor technology, to measure blood parameters such as hemoglobin levels non-invasively. The collected data are then analyzed using machine learning algorithms to accurately predict various blood counts, including HCT, RBCs, MCV, MCH, and MCHC. This approach not only enhances diagnostic efficiency for conditions like Thalassemia, which typically require frequent blood testing, but also offers a quicker, cost-effective, and patient-friendly alternative to traditional blood sampling methods, thereby eliminating the associated discomfort and potential complications [11].

Xu et. al. in 2019 introduced the Simulated Annealing Extreme Learning Machine (SAELM), a hybrid machine learning technique designed to enhance the prediction of Thalassemia, a severe hereditary blood disorder where early detection is critical due to its incurable nature. The SAELM algorithm merges the rapid computation and superior generalization abilities of the Extreme Learning Machine (ELM) with the robust optimization capabilities of Simulated Annealing

(SA). This innovative combination aims to optimize the initialization of weights and biases in ELM, addressing its inherent limitations and enhancing its predictive accuracy. The results of the study demonstrate that SAELM significantly outperforms traditional ELM across several key performance metrics, highlighting its potential as an effective medical diagnostic tool for Thalassemia screening[12].

Akhtar et. al. in 2020, utilized machine learning to enhance the prognosis process for thalassemia by analyzing complete blood count (CBC) data. This research marks the first attempt to apply Linear Discriminant Analysis (LDA) to CBC parameters to accurately predict thalassemia, addressing the need for efficient diagnostic methodologies. Parameters such as WBC, RBC, HB, HCT, Platelets, and Ferritin were analyzed, with RBC, HB, and Ferritin identified as particularly critical in predicting thalassemia effectively. This approach offers a potential pathway to replace more invasive, costly, and time-consuming diagnostic methods, aiming to streamline and improve the accuracy of thalassemia diagnostics through data-driven techniques[13].

Sadiq et al. in 2021, explore ensemble machine learning models to identify β-Thalassemia carriers using red blood cell indices. Their research develops a Voting Classifier, named SGR-VC, which combines Support Vector Machine (SVM), Gradient Boosting Machine (GBM), and Random Forest (RF) to enhance detection accuracy. Using data from 5,066 individuals, the ensemble achieves a classification accuracy of 93%, demonstrating the efficacy of integrating multiple algorithms. This approach not only improves diagnostic accuracy but also offers a cost-effective tool for early screening and management of β-Thalassemia, outperforming individual models in precision, recall, and F1-score [14].

Laeli et al. in 2020, highlighted the impact of hyperparameter optimization in SVMs on thalassemia classification, utilizing a dataset from Harapan Kita Children and Women's Hospital in Jakarta, which comprises 150 samples with 11 features. By employing Grid Search to fine-tune the C and gamma parameters of an SVM with an RBF kernel, the research achieved significant enhancements in SVM performance for thalassemia classification. The results indicate that optimal hyperparameters can substantially increase accuracy, reaching 100% in some instances. This demonstrated the potential of hyperparameter optimization to significantly improve the efficacy of machine learning models in medical diagnostics, particularly for thalassemia [15].

Purwar et al., 2021 presented a novel approach to diagnosing thalassemia by combining deep learning with clinical data analysis. The study Introduced a deep convolutional neural network (CNN) model that analyzes both clinical features from blood tests and morphological features from blood smear images. Principal component analysis (PCA) is used to reduce feature dimensionality and computational complexity. The study employed machine learning algorithms like Naive Bayes, Random Forest, and KNN, achieving high classification accuracy of 99±1%, with specificity and sensitivity rates at 100% [16]. Tressa et al. (2023) explored the application of machine learning algorithms to classify Alpha Thalassemia in patients based on genetic mutations. Alpha Thalassemia is a genetic blood disorder that affects hemoglobin production. The study uses a data-driven approach, utilizing patient records including demographic data, health history, and lab results, applying supervised learning techniques to identify patterns indicative

of the disorder. The primary algorithms utilized are Decision Trees, Artificial Neural Networks, Naive Bayes, and Support Vector Machines. The classifier achieves a high accuracy rate of 95% and a Kappa statistic of 0.947, showcasing its potential to enhance diagnosis and treatment strategies for Alpha Thalassemia [17].

Abdulhay et al. in 2021 presented a method to diagnose and differentiate between various blood disorders using convolutional neural networks (CNNs). This study leverages high-resolution images of blood samples to train a CNN, bypassing traditional blood tests like CBC. The CNN, designed using Python, achieves an overall testing accuracy of 93.4%, offering a promising, low-cost, and fast diagnostic alternative that doesn't require a lab setting [18].

Meti et al. in 2023 explored the use of various machine learning (ML) models to enhance the screening and diagnosis processes for α-thalassemia, assessing several algorithms including Logistic Regression, Decision Tree, XGBoost, Random Forest, and LightGBM. Decision Trees emerged as the most accurate, with an 87% success rate. The study also integrates explainable AI methods, notably SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), to demystify the model's decisions for medical professionals. This strategy not only boosts diagnostic precision but also increases trust and understanding of ML outputs within the healthcare community[19].the paper Saleem et al. (2023) analyzes various feature selection techniques to enhance the accuracy of predicting thalassemia. It employs methods such as Chi-Square, Exploratory Factor Score, Recursive Feature Elimination, and others to identify the most significant features for thalassemia prediction. Multiple classifiers, including K-Nearest Neighbors, Decision Trees, and Gradient Boosting, were tested, with the Gradient Boosting Classifier achieving a top accuracy of 93.46%. This study showcases the potential to improve diagnostic models for thalassemia through sophisticated feature selection and machine learning strategies[20].

Ip et al. in 2023 discussed the integration of AI technologies in the field of hematology to enhance diagnostic accuracy and efficiency. The review highlighted several AI-assisted methods and their application in diagnosing various hematologic disorders, including thalassemia. It points out the potential of AI to improve diagnostic workflows, reduce errors, and predict disease outcomes. However, it also acknowledges several limitations such as the need for extensive data sets for AI training, the possibility of systematic errors and bias in AI algorithms, and concerns over data privacy[21]. Phirom et al., (2022) introduced and evaluated a machine learning (ML) framework called DeepThal, designed to predict α+-thalassemia trait using red blood cell indices from a retrospective study of 594 subjects. They utilized various ML models, including convolutional neural networks (CNNs), and demonstrated that DeepThal significantly outperformed other models and traditional diagnostic methods, achieving an accuracy of 80.77%, sensitivity of 70.59%, and specificity of 81%. The study underscores the potential of ML to enhance the diagnosis of α+-thalassemia trait and support widespread screening efforts, especially in areas where the disease is prevalent [22].

Fu et al. in 2021 focused on developing a machine-learning-based classifier using Support Vector Machine (SVM) algorithms to enhance the diagnosis of thalassemia compared to non-thalassemia anemias in Taiwanese adult patients. By

analyzing complete blood count parameters, the classifier distinguishes thalassemia from other microcytic anemias, such as iron deficiency anemia (IDA) and anemia of inflammation (AI). Utilizing retrospective data from 350 patients and applying SVM with Monte-Carlo cross-validation, the classifier achieved a notable improvement in diagnostic accuracy, evidenced by an average AUC (Area Under the Curve) of 0.76 and an error rate of 0.26, outperforming traditional diagnostic indices for differentiating between thalassemia and IDA [4].

Zhang et al. in 2023 paper discussed the TT@MHA tool, a machine learning (ML) algorithm crafted to differentiate thalassemia trait (TT) from iron deficiency anemia (IDA) in patients with microcytic hypochromic anemia (MHA). The study analyzed retrospective data from 798 MHA patients using five ML models: Linear SVC (L-SVC), Support Vector Machine (SVM), Extreme Gradient Boosting (XGB), Logistic Regression (LR), and Random Forest (RF). These models were evaluated against six established discriminant formulas. The RF model emerged as the most effective, demonstrating high sensitivity (91.91%), specificity (91.00%), accuracy (91.53%), and an AUC of 0.942. To support healthcare providers, particularly in rural areas with limited technological resources, a webpage tool for the TT@MHA model was developed [23].

Das et al. in 2022 study assessed various machine learning algorithms (MLAs) and discriminant formulas for screening β-thalassemia trait (BTT) among Indian antenatal women. It involved testing 13 MLAs and 27 discriminant formulas on a dataset of 2,942 antenatal females to evaluate their effectiveness in distinguishing BTT from other types of microcytic anemia. Among the MLAs examined were Random Forest (RF), Extreme Learning Machine (ELM), Gradient Boosting Classifier (GBC), and Logistic Regression (LR). These algorithms were evaluated based on their sensitivity, specificity, Youden's Index, and Area Under the Curve (AUC-ROC). The ELM and GBC algorithms, in particular, stood out for their superior performance in terms of Youden's Index and AUC-ROC [7]. The Çil et al., (2020) article outlined the creation of a decision support system that employs Extreme Learning Machine (ELM) and Regularized Extreme Learning Machine (RELM) algorithms to distinguish between β-thalassemia and iron deficiency anemia (IDA) using complete blood count (CBC) parameters. The study included 342 patients and aimed to provide high accuracy and performance while reducing computational costs and complexity compared to traditional methods. The performance metrics were impressive, with RELM achieving an accuracy of 95.59% in scenarios involving both male and female patients, and ELM excelling with female patients at an accuracy of 96.30%. This system addresses the challenge of differentiating between β-thalassemia and IDA, which often exhibit similar symptoms and CBC indices, by offering a cost-effective and efficient diagnostic tool [8].

Ayyıldız & Arslan Tuncer in 2020 explored the use of machine learning (ML) techniques and Neighborhood Component Analysis (NCA) feature selection to differentiate between iron deficiency anemia (IDA) and beta thalassemia (β-thalassemia) using red blood cell (RBC) indices. The study utilized data from 342 patients, employing algorithms like Support Vector Machine (SVM) and K-Nearest Neighbor (KNN), and achieved a 97% Area Under the ROC curve (AUC), indicating a high level of predictive accuracy [24].

Lee et al. in 2021 study detailed the creation and evaluation of a CNN-based AI algorithm aimed at detecting Hemoglobin H (HbH) inclusions in blood smears, a method that promises to enhance the detection rate, efficiency, and testing quality for alpha-thalassemia carriers and HbH disease. This approach modernizes the traditional, labor-intensive microscopic analysis by utilizing digital images of HbH-positive and HbH-negative blood smears, captured under various magnifications and across different scanning platforms. The algorithm demonstrated high sensitivity (approximately 91%) and specificity (99%) at 100x magnification. Moreover, it proved effective at lower magnifications (40x and 60x) and maintained consistent performance across diverse imaging systems, underscoring its robustness and adaptability for clinical use [25].

Diaz-del-Pino et al. in 2023 study presented a neural network-based AI model designed to aid clinicians in diagnosing various hematological diseases through routine blood count tests. Achieving up to 96% accuracy in binary classification tasks, the model is benchmarked against traditional machine learning algorithms, such as gradient boosting decision trees. Utilizing 4,124 hemograms from Hospital Clínico San Carlos in Madrid, Spain, the researchers employed advanced data preprocessing and feature engineering techniques to optimize model performance. Additionally, they conducted extensive data processing and integrated neural networks with traditional machine learning methods to assess the effectiveness of their model. A significant aspect of their approach was the application of contribution analysis techniques, which helped interpret the AI model's decision-making process, thereby increasing the transparency and understandability of AI decisions in clinical settings [26].

Feng et al. in 2022 focused on the development of a machine learning model using random forest to improve the screening of α-thalassemia carriers from patients with low Hemoglobin A2 (HbA2) levels. The study utilized data from 1,613 patients and employed 14 machine learning algorithms to optimize the screening process. The random forest model, selected for its superior performance, significantly enhanced the positive predictive value (PPV) and other metrics compared to traditional hemoglobin electrophoresis (HE) [27].

Basu et al. in 2022 study demonstrated the use of machine learning techniques like K-means clustering and XGBoost to assess and categorize the severity of β-thalassemia based on oxidative stress biomarkers and other biochemical parameters. By combining multiple ML approaches, the study achieves high diagnostic accuracy and enhances treatment specificity, showing potential for significant impacts on clinical practice by providing reliable disease severity assessments and predicting key biomarkers from accessible clinical data[28].

Mo et al. in 2023 details the development of a deep neural network (DNN) aimed at improving thalassemia screening using red blood cell (RBC) indices, marking a significant advancement over traditional statistical methods. The study demonstrated the potential of machine learning techniques, particularly DNNs, to enhance existing diagnostic models significantly, focusing on the efficiency and accuracy of thalasemia screening protocols. By incorporating diverse features such as age and red cell distribution width (RDW) into the model, the accuracy is not only enhanced but also highlighted the complexity of thalassemia as a condition influenced by multiple physiological parameters. This innovative approach could

lead to more personalized and accurate diagnostic techniques for hematological disorders, setting a new standard for medical diagnostics in the field [29].

Karollus et al. in 2021 presented a deep learning model that predicts ribosome load from mRNA sequences, useful for analyzing genetic variants in clinical settings. It demonstrated the model's application by identifying a mutation in the HBB gene's 5'UTR associated with beta-thalassemia, highlighting its potential in diagnosing and understanding thalassemia through crucial genetic insights [30].

Laengsri et al. in 2019 Introduced ThalPred is a web tool that uses machine learning to distinguish between thalassemia trait and iron deficiency anemia more effectively. Employing algorithms like SVM, it outperforms traditional methods in accuracy and reliability. Its user-friendly interface makes it a practical choice for healthcare providers to enhance anemia screening in clinical settings [31].

Zhang et al. in 2022 presented a new diagnostic approach using MALDI-TOF mass spectrometry to improve the rapid screening of thalassemia. This study utilized a machine learning model to analyze haemoglobin chain data from 674 samples to discriminate thalassemia patients from controls. The logistic regression model showed outstanding performance with an AUC of 0.99, demonstrating high diagnostic accuracy [32].

Tran et al. in 2023 explored the development and application of both expert and AI-based clinical decision support systems (CDSS) for thalassemia screening in the Vietnamese population. The study included 10,112 medical records and utilized machine learning models to improve prenatal screening, achieving high accuracy rates in identifying thalassemia carriers, The study demonstrated the effective use of CDSS, both expert and AI-based, in a clinical setting to enhance the accuracy and efficiency of prenatal screening for thalassemia, highlighting its potential for broader application in healthcare systems[33].

Rodríguez-González et al. in 2023 delves into the development of a machine learning model using an extreme learning machine (ELM) algorithm to enhance the diagnosis of different types of anemia, including beta thalassemia trait (BTT), iron deficiency anemia (IDA), and hemoglobin E (HbE). The study utilized historical laboratory data to train the model and demonstrated high performance metrics, including an accuracy of 99.21%, sensitivity of 98.44%, and precision of 99.30% [34].

Y. Zhang et al. in 2019 explored the use of machine learning algorithms to develop a predictive model for identifying inhibitors against K562 cells, which are used in the study of β-thalassemia [35].

Badat, M. et al. in 2023 demonstrated how machine learning models can predict and mitigate off-target mutations, thereby enhancing the safety and efficacy of gene editing. Researchers utilized adenine base editors to efficiently correct the HbE mutation in hematopoietic stem cells, showing potential in reducing the necessity for lifelong blood transfusions and minimizing risks such as insertional mutagenesis. This approach highlighted the promising role of machine learning in advancing gene therapy for complex genetic disorders [36].

Rustam et al. in 2022 presented a sophisticated approach to enhancing the screening of β-thalassemia carriers through the use of machine learning (ML) models on red blood cell indices from Complete Blood Count (CBC). This study

specifically tackles the challenges of data imbalance and feature selection, employing methods like the Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic (ADASYN) for oversampling, alongside Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) for feature reduction. The research demonstrated that their ML approach, which extensively tests various algorithms including Decision Trees, Gradient Boosting Machine, and Support Vector Classifier, significantly improves the accuracy of β-thalassemia carrier detection [37].

Uçucu & Azik in 2024 investigates the use of artificial intelligence (AI), particularly artificial neural networks (ANNs) and decision trees, to differentiate between β-thalassemia minor (BTM) and iron deficiency anemia (IDA) using complete blood count (CBC) data. This study aims to create an efficient, cost-effective model that improves upon traditional discriminant indices and diagnostic methods[38].

Sani et al. 2024 discussed the widespread issue of hemoglobinopathies, such as thalassemia and other structural hemoglobin variants, emphasizing their significant impact on global health. The paper reviews recent advancements in clinical analytical techniques and the integration of artificial intelligence in the detection and research of these conditions, pointing out the lack of comprehensive reviews in this field. Key diagnostic technologies like high-performance liquid chromatography, capillary zone electrophoresis, and mass spectrometry are enhanced by AI applications, including machine learning models and portable point-of-care tests. The article also covers specialized genetic techniques for identifying and validating unknown or novel hemoglobins, stressing the importance of improving these technologies to manage hemoglobinopathies effectively [39].

Ibrahim et al. (2024) presented a late fusion-based machine learning model designed to predict β-thalassemia carriers efficiently. The study leverages four distinct machine learning algorithms—logistic regression, Naïve Bayes, decision trees, and neural networks—achieving individual accuracies of 94.01%, 93.15%, 97.93%, and 98.07% respectively, using a feature-based dataset. The late fusion model, which integrates the outcomes of these algorithms through a fuzzy logic system, demonstrated an overall accuracy of 96%. This model outperforms previous methods in terms of efficiency, reliability, and precision, suggesting significant potential for improving early diagnosis and management of β-thalassemia carriers [40].

Long & Bai, 2024 study analyzed 7,621 cases from Jiangjin District, Chongqing, China, focusing on blood routine indicators such as mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), red blood cell count (RBC), and mean corpuscular hemoglobin concentration (MCHC). The least absolute shrinkage and selection operator (LASSO) regression was employed to select these indicators for their high predictive value. The model achieved an area under the ROC curve (AUC) of 0.911, indicating a strong predictive ability. The study highlighted the effectiveness of using routine blood test indicators combined with machine learning to predict thalassemia, offering a faster and more cost-effective approach than traditional genetic testing [41].

Jahan et al. in 2021 assessed the use of red cell indices and machine learning algorithms for beta thalassemia trait (BTT) screening among antenatal women. Conducted as a cross-sectional study at a tertiary care hospital, it tested C4.5 and Naive Bayes classifiers, along with an artificial neural network (ANN). Findings revealed that while individual red cell indices were inadequate for effective screening, the integrated ANN model achieved an accuracy of 85.95%, with sensitivity and specificity at 83.81% and 88.10% respectively. These results indicate potential for effective use of these models in peripheral settings for thalassemia screening [42].

Setiawan et al. in 2021 explored the development of a fuzzy-based model for predicting various types of thalassemia (major, intermedia, minor, and not thalassemia) in children using complete blood count (CBC) data. This novel model employs fuzzy logic to handle the uncertainty and variability inherent in medical diagnosis, offering a refined approach by distinguishing between four categories of thalassemia, compared to previous models that identified three. The study highlighted the model's successful application in distinguishing thalassemia types, validated against pediatrician diagnoses with CBC data. The fuzzy-based model was implemented in software, which demonstrated high concordance with expert opinion in testing scenarios [43].

Setiawan et al. in 2020 discussed the application of the Random Forest (RF) algorithm to classify thalassemia data from Harapan Kita Children and Women's Hospital in Indonesia. The study uses a dataset comprising 150 patients, with 82 diagnosed with thalassemia and 68 as non-thalassemia. The RF model was trained with various proportions of the data, ranging from 50% to 85%, achieving high classification metrics, with the best results showing 100% accuracy, precision, and recall when trained with 70% to 85% of the data. This model offers a robust tool for early detection and classification of thalassemia, potentially enhancing patient management and outcomes [44]. Uçucu et al., (2022) investigates the use of various machine learning models, including Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), Naive Bayes, and Decision Trees, to predict hemoglobin variants such as HbS and HbD Los Angeles carriers. This study utilized a dataset of 238 observations to train these models, with features including age, sex, various blood count and hemoglobin metrics, and retention times from high-performance liquid chromatography (HPLC). The models were assessed using 7-fold cross-validation. The study highlighted the effectiveness of the deep learning model, which excelled with an accuracy, specificity, sensitivity, and F1 score of 0.99, indicating its potential utility in clinical settings for hemoglobinopathy detection [45].

Y. Setiawan et al. in (2024) Introduced a hybrid machine learning model integrating Neuro-SVM (Neural Networks and Support Vector Machines) to predict treatment outcomes in beta-thalassemia patients who also have Hepatitis C. The model showed high accuracy rates of 98.83% in group 1 and 99.75% in group 2, indicating excellent potential for clinical decision support. This research is critical as it could help identify patients who would benefit from direct anti-viral agents (DAAs), thus optimizing treatment strategies [46]. develops a decision tree model for the early detection of Thalassemia Major using ID3, C4.5, and CART algorithms. Data was collected through interviews and medical records from a hospital in

Surabaya, Indonesia. The C4.5 algorithm outperformed others with a 100% accuracy rate, demonstrating no signs of overfitting or underfitting. It also conducted automatic feature selection, enhancing the model's efficiency and interpretability. The model's ability to use simple Yes/No data effectively reduces complexity in diagnosing Thalassemia Major [47].

Liu & Liu in 2024 examined two machine learning strategies, Principal Component Analysis combined with Logistic Regression (PCA-LR) and Partial Least Squares Regression (PLS), used to manage high dimensionality and multicollinearity in clinical data. It finds a higher prediction accuracy for PLS (92.5%) compared to PCA-LR (87.5%) and discussed challenges like selecting principal components and regularization parameters. The study underscores the value of dimensionality reduction in handling large-scale clinical data and suggests further investigation into these techniques for various stages of Thalassemia[6].

Ferih et al. in 2023 reviews various machine learning algorithms used in the diagnosis and differentiation of thalassemia from other forms of microcytic anemia. The paper emphasizes the role of artificial intelligence (AI) in enhancing diagnostic accuracy, reducing unnecessary tests, and aiding in the management of thalassemia. The study highlighted several AI techniques including k-nearest neighbor (k-NN), Naïve Bayesian, decision trees, and neural networks, all of which show promise in distinguishing thalassemia based on complete blood count (CBC) parameters [5].

1. Technological Advancements and Diagnostic Efficacy

One of the paramount strengths highlighted in the review is the utilization of diverse ML algorithms ranging from AdaBoost to deep learning models like convolutional neural networks (CNNs). For instance, AdaBoost achieved a remarkable 100% accuracy in one study [9], underscoring the potential of ensemble methods in improving diagnostic precision. Similarly, the integration of deep learning techniques, particularly through high-resolution medical imaging and CNNs, has enhanced the ability to discern subtle morphological changes associated with Thalassemia[10]. These advancements not only augment the diagnostic process but also significantly reduce the reliance on invasive traditional methods, making diagnosis quicker and less cumbersome for patients.

2. Challenges in Data Diversity and Model Transparency

Despite these advancements, the review consistently points to challenges related to data diversity and model transparency. The efficacy of ML models heavily depends on the diversity and volume of the dataset on which they are trained. Several studies noted limitations due to small sample sizes or the lack of comprehensive demographic representation, which could impact the generalizability and applicability of these algorithms across different populations [12], [22]. Furthermore, the need for transparent, interpretable models is critical, as medical practitioners must understand and trust the machine learning outputs to integrate them effectively into clinical workflows. Studies like those by Meti et al. (2023) and Saleem et al. (2023) emphasize the integration of explainable AI techniques, which help demystify ML decisions and thus foster trust among healthcare providers[19], [20].

3. Potential for Broader Application and Future Research

The review highlighted a promising trajectory for the application of ML in Thalassemia diagnosis that could be extended to other genetic and hematological disorders. The future of ML in hematology appears to hinge on overcoming current limitations through innovations in data collection, model training, and integration into existing healthcare systems. Further research is suggested to focus on creating larger, more diverse datasets and developing models that are not only accurate but also adaptable to various clinical environments across the globe.

Table 1. Overview of Literature on Thalassemia Prediction Using AI

Ref	Algorithms	Datasets Utilized	Advantages	Disadvantages	Accuracy
[9]	KNN, Logistic Regression, SVM, Naïve Bayes, Random Forest, AdaBoost, XGBoost, Decision Tree, MLP, Gradient Boosting	Processed dataset	Utilizes a variety of ML models to optimize prediction accuracy; comprehensive use of clinical and genetic data	Requires complex data preprocessing; high variance in model performance based on data quality and selection	AdaBoost: 100%
[10]	Deep Learning, U-Net architectur, (CNNs), Prediction Time Augmentation (PTA)	Medical imaging datasets including MRI and CT scans of patients diagnosed with various types of Thalassemia.	Non-invasive, provides quick and accurate diagnosis, reduces dependency on invasive blood tests.	Requires high-quality imaging data, needs substantial training data to achieve high accuracy.	94% accuracy in identifying thalassemia types
[11]	Linear Regression, Decision Trees	Photoplethysmography data and over 800 CBC reports from pathological labs	Non-invasive, pain-free, no blood sample required, potentially high patient compliance	Limited by the specificity and sensitivity of the optoelectronic sensors, may need extensive validation for clinical use	Various accuracies for different blood parameters, with significant predictions for anemia classification
[12]	Simulated Annealing Extreme Learning	Thalassemia dataset	Can adapt learning rates and parameters dynamically, reducing the need for manual	May require more computational resources than standard ELM due to	90.12% for SAELM 88.76% for ELM

	Machine (SAELM), Extreme Learning Machine (ELM)		tuning.	adaptive mechanisms.	
[13]	Linear Discriminant Analysis (LDA)	CBC Clinical data including complete blood counts and iron profiles from thalassemia patients	Effective in distinguishing between different types of thalassemia based on key blood parameters	May not account for all variability in smaller or less homogeneous datasets	Accuracy not directly stated; significant variables identified include RBC, HB, Ferritin
[14]	SVM, GBM, RF, Voting Classifier (SGR-VC)	Dataset from Punjab Thalassemia Prevention Programme comprising 5066 individuals' blood indices	High accuracy, robustness against overfitting due to ensemble approach, good generalization	Can be computationally intensive, requires tuning of multiple models	RF: 93% SVM:90% GBM:91% SGR-VC:93%
[15]	SVM with RBF kernel, Grid Search for hyperparameter optimization	Thalassemia data from Harapan Kita Children and Women's Hospital, Jakarta; 150 samples, 11 features	Optimizes SVM performance by systematically searching through a range of hyperparameters; helps in finding the model with the best generalization on unseen data	Computationally intensive, can be time-consuming especially with large datasets and extensive hyperparameter grids	ACC: 100% with optimal parameters (C = 428.13, gamma = 0.0000183)
[17]	Support Vector Machine (SVM), Decision Trees, Random Forest	Genetic data from a biobank including over 10,000 subjects with detailed genotyping	High accuracy and interpretability, ability to handle large and complex genetic data	Requires extensive computational resources, potential overfitting due to high model complexit	ACC: 95%,

[18]	Convolutional Neural Network (CNN)	High-resolution blood sample images from multiple datasets	Fast and low-cost diagnosis without the need for a laboratory; can differentiate between multiple diseases using blood sample images	Potential for overfitting; relies heavily on image quality and quantity for training	ACC: 93.4%
[19]	Explainable AI (XAI) Logistic Regression, Decision Tree, XGBoost, Random Forest, LightGBM	Genetic data from population studies including gene expression profiles and variant analysis	Enhances transparency and interpretability of ML predictions, fostering trust among clinicians	Complexity of models can limit user understanding despite explainability efforts	92% accuracy
[20]	Feature selection techniques applied on logistic regression models	Data from thalassemia patients including blood indices like RBC count, MCV, MCH	Improved model performance by reducing overfitting and enhancing interpretability	Potential loss of important information through feature reduction, complexity in feature selection process	Overall accuracy of 87.5%
[21]	ML, DL (specific models: MLP, RNN, CNN); Specific methods: MRMR-XGB, ANOVA-MLP, RFE-KNN, CellaVision, Morphogo	Various hematologic datasets Blood cell and bone marrow images, genetic data, cytogenetics	Enhanced diagnostic accuracy, improved detection of hematologic disorders, rapid processing	High data requirements, potential bias, limited by data quality and training	SVM best performance with an accuracy 87.7%
[22]	Deep Neural Networks (DNNs) CNNs (DeepThal), SVM, MLP	Data from 23,000 patients with red blood cell indices	Provides a scalable model for large datasets, automates and enhances the prediction of α+-thalassemia using routine blood test data	Requires substantial computational resources, potential for overfitting on extensive data	Predictive accuracy of 95%
[4]	Support Vector Machine (SVM)	Retrospective analysis, 350 patients	Enhances the accuracy of thalassemia screening, reduces false diagnosis rates, automated process reduces human error	Requires extensive training data, potential bias in machine learning models, model specifics and techniques not detailed	AUC: 0.76

[6]	PCA-LR (Principal Component Analysis followed by Logistic Regression) ,PLS (Partial Least Squares)	Clinical records from a hospital in the Guangxi Zhuang Autonomous Region, China (60 individuals, 110 genes each)	Effective handling of multicollinearity, small sample sizes; high accuracy with PLS	Higher complexity and computation required for PLS compared to simpler models	PCA-LR: 87.5%, PLS: 92.5%
[5]	k-NN, Naïve Bayesian, Decision Trees, Neural Networks	Various datasets including patient data for Thalassemia and Iron Deficiency Anemia (IDA) across multiple studies	AI significantly aids in diagnosing and differentiating Thalassemia from similar conditions, improving speed and reducing unnecessary testing.	Dependence on the quality and size of datasets, which can affect the generalizability and accuracy of the models..	Varies by model; up to 98.7% specificity and sensitivity
[7]	RF, ELM, GBC, LR, among others	Dataset of 2942 antenatal females from PGIMER, Chandigarh	Comprehensive approach combining multiple algorithms to enhance screening accuracy.	Complexity of managing and interpreting results from multiple algorithms; potential for overfitting	ELM and GBC showed highest AUC-ROC 0.92 (GBC and ELM)
[8]	Extreme Learning Machine (ELM), Regularized Extreme Learning Machine (RELM)	342 patients	Allows for rapid and accurate classification of anemia types, reducing diagnostic time and costs.	Potentially less effective with smaller datasets or underrepresented conditions in the training data.	95.59% (RELM), 96.30% (ELM)
[23]	L-SVC, SVM, XGB, LR, RF (TT@MHA)	Retrospective data, 798 patients	Provides a user-friendly web interface for rapid screening; reduces the need for invasive procedures	Dependent on the quality and variability of the input data; limited external validation	AUC: 0.942
[24]	SVM, KNN, NCA feature selection	342 patients	Enhances model accuracy by focusing on the most informative features; reduces	Can potentially overlook important but less obvious features; may be sensitive to noise in the data	AUC: 97%

			computational load		
[25]	Convolutional Neural Network (CNN)	Digital images of HbH-positive and HbH-negative blood smears	High accuracy and efficiency, reduces labor and time, minimizes human error	Dependent on the quality of digital images, requires precise staining and imaging conditions	Sensitivity: ~91%, Specificity: ~99% overall accuracy of 97.6%
[26]	Neural Networks, SVM, Random Forest, Gradient Boosting Decision Trees	114,789 hemograms from the Hospital Clínico San Carlos (Madrid, Spain)	High accuracy and improved efficiency in disease diagnosis using only routine blood count tests; use of contribution analysis for model interpretability.	High dependence on image quality and proper staining; requires extensive training data; potential for algorithmic bias	Up to 96.4% accuracy for binary classification of hemoglobinopathies
[27]	Random Forest	Red blood cell parameters from a database of individuals with low HbA2 levels indicative of alpha-thalassemia carriers. 1,613 patients	Allows for effective discrimination of alpha-thalassemia in cases with typically challenging low HbA2 levels, making use of accessible clinical data.	The model's effectiveness may be limited by the specificity and variability of HbA2 levels in the tested population	0.915
[28]	K-means, Random Forest, XGBoost, Decision Tree, Neural Networks	Oxidative stress biomarkers, hormonal and ferritin levels from 105 patients	Effective in identifying carriers with low HbA2 levels, which are typically difficult to screen; automates and simplifies the screening process.	Limited to cases with low HbA2 levels; effectiveness in broader populations not established.	XGBoost: 100% after K-cross validation
[29]	Deep Neural Networks (DNNs)	8,693 records including genetic tests and 11 features	Uses comprehensive blood cell data to train the DNN, potentially increasing the predictive accuracy for thalassemia.	High computational costs; requires extensive data preprocessing and could be prone to overfitting if not properly regulated.	0.897

[30]	Deep Neural Networks with Frame Pooling	Over 200,000 5'UTR sequences ranging 25 to 100 nucleotides	Enables detailed and accurate prediction of ribosomal load across different 5'UTR lengths; useful for understanding gene expression.	High computational demand; requires substantial training data to achieve high accuracy.	Pearson correlation up to 0.964 on MPRA data
[31]	k-NN, DT, RF, ANN, SVM	186 patients (146 TT, 40 IDA)	Utilizes a variety of ML algorithms to increase the robustness and accuracy of predictions; accessible as a web-based tool	Limited to the specificity of the dataset which might not generalize to other populations or settings .	External accuracy: 95.59%, AUC: 0.98
[32]	MALDI-TOF Mass Spectrometry	674 samples including thalassemia and control groups	Rapid screening capability, high-throughput analysis, minimal sample preparation.	Potential for interference in complex samples, high equipment cost, requires skilled operation	AUC: 0.99
[33]	Expert System, AI-based CDSS (MLP, SVM, RF, KNN)	Data from 10112 medical records of first-time pregnant women and their husbands, including 1992 cases for training	High accuracy in predicting thalassemia carriers, leveraging AI to assist in medical decision-making.	Lack of transparency in how AI models make predictions ("black box" issue), which can affect trust and reliability in clinical settings.	Expert system: 98.45%, MLP: 97% (general), 97.81% (women only)
[34]	Extreme Learning Machine (ELM)	190 data samples for BTT, IDA, HbE, and combination anemias	Utilizes rapid and efficient ELM algorithm, suitable for large-scale analysis, high accuracy and precision in diagnosing anemia types.	Requires quality data for training, limited to conditions present in dataset, high dependency on dataset quality and preprocessing	99.21% accuracy, 98.44% sensitivity , 99.30% precision, and 98.84% F1 score.
[35]	Adaboost, Bayes Net, Random Forest, Random Tree, C4.5, SVM, KNN, Bagging	Drug screening datasets, including compound activity data against K562 cells 117 inhibitors, 190non-	High predictive accuracy for identifying potential inhibitors, useful for drug discovery and development.	Requires large and diverse chemical datasets to train effectively; may not generalize to other cell lines without retraining. further validation needed	85%

		inhibitors			
[36]	Machine Learning (DeepHaem model for predicting off-target effects) (ELM)	Patient-derived HSCs, xenograft assays in NSG mice	ELM can handle large datasets efficiently and effectively with high speed in training and prediction.	As with many AI models, there can be challenges in interpreting the model's decision-making process, leading to potential issues with trust among clinicians.	High editing efficiency (up to 98.8%)
[37]	SMOTE, ADASYN, PCA, SVD, Decision Tree, Gradient Boosting Machine, AdaBoost, Support Vector Classifier, Random Forest	Dataset from Punjab Thalassaemia Prevention Programme, including CBC data of 5066 individuals	High accuracy, addresses data imbalance with SMOTE and ADASYN, employs PCA and SVD for feature reduction	Complexity of models might increase computational cost, potential "black box" issue with deep learning models ,Limited to the dataset from a specific geographic location, potential overfitting concerns	Up to 0.96
[38]	Artificial Neural Networks, Decision Trees	396 individuals (216 IDA, 180 BTM)	High accuracy in differentiating IDA and BTM using CBC data alone. Fast and inexpensive.Can handle nonlinear relationships and model complexities.	Limited by its retrospective data collection and lack of genotype reports, potentially affecting the precision and accuracy of its AI models.	ANN: 99.5%
[39]	Machine learning models integrated with smartphone-based microscopic classification and complete blood counts	Clinical screening and research data, gene-based technologies, molecular diagnosis datasets	Integratable into medical lab protocols, supports early disease detection and prognosis	High complexity in integration and interpretation, requires substantial computational resources	Reliable detection of malignant cell populations, support for chromosome banding analysis
[40]	Logistic Regression, Naïve Bayes, Decision Trees, Neural Networks, Fuzzy Logic	Features-based dataset from Internet of Medical Things (IoMT) enabled devices	Higher efficiency, reliability, and precision compared to previous models. Uses a fusion of multiple machine learning algorithms.	lacks detailed demographic information on the dataset.	Overall model: 96% Logistic Regression: 94.01% Naïve Bayes: 93.15% Decision Tree:

					97.93% Neural Network: 98.07%
[41]	Least Absolute Shrinkage and Selection Operator (LASSO) regression	7,621 cases (847 thalassemia patients, 6,774 non-thalassemia) from the Jiangjin area of Chongqing, 2018-2022	The model effectively predicts thalassemia using routine blood tests, reducing the need for expensive genetic tests. Provides an economical and rapid screening method for early diagnosis of thalassemia in pregnancy.	The model may not accurately predict all types of thalassemia, particularly b-thalassemia outside Asia. It also doesn't distinguish between varying degrees of thalassemia severity or different genotypes like iron-deficiency anemia.	Overall AUC: 0.911; MCV: 0.907, MCH: 0.906, RBC: 0.796, MCHC: 0.795
[42]	C4.5 Classifier, Naive Bayes Classifier, Artificial Neural Network (ANN)	3947 antenatal women undergoing thalassemia screening at a tertiary care hospital	High accuracy in detection of BTT using combined red cell indices, good sensitivity (ANN: 83.81%) and specificity (ANN: 88.10%)	Performance dependent on composition of non-BTT group, lower accuracy compared to studies excluding other anemia types	C4.5: 88.56%, NB: 82.49%, ANN: 85.95%
[43]	Fuzzy Logic Model	CBC data from pediatric patients at Abdoel Moeloek Hospital, Lampung Province Moeloek Hospital, Lampung Province	Offers rapid and economical prediction of thalassemia types using readily available CBC data. Can differentiate between major, intermedia, minor, and non-thalassemia cases.	Accuracy not quantitatively assessed. Implementation and validation in broader clinical settings not discussed. Limited to data from one hospital, which may not represent wider population diversity.	Not specified
[44]	Random Forest	150 patient data from Harapan Kita Children and Women's Hospital, Indonesia (82 thalassemia, 68 non-	High accuracy, precision, and recall. Can handle missing data and provides estimates of variable importance. Suitable for large datasets with high dimensionality.	Requires large amounts of data for training to achieve high accuracy. Can be computationally intensive due to multiple decision trees.	100% (70-85% training data range)

		thalassemia)			
[45]	ANN, KNN, Naive Bayes, Decision Trees	238 observations from patients suspected of carrying HbD or HbS variants, including comprehensive blood count and HPLC data	High accuracy across models, especially with deep learning. Demonstrated effective use of machine learning for hemoglobin variant detection in medical diagnostics.	Dependence on specific data features such as RT for optimal performance. Naive Bayes showed lower performance compared to other models.	Up to 0.99 (Deep Learning Model without RT)
[46]	Neuro-SVM (Neural Networks and Support Vector Machines) Hybrid Neuro-SVM, MLP, SVM, NB	341 β-TM patients infected with HCV genotype 4, from different centers in Cairo and Upper Egypt	The hybrid model combines SVM and ANN, enhancing predictive accuracy. It offers a robust tool for clinicians to predict treatment response efficiently.	Complex model requiring extensive computational resources. Model validation limited to specific patient demographics in Egypt.	Hybrid Neuro-SVM: 98.8% in Group 1; 99.87% in Group 2
[47]	ID3, C4.5, CART	Data from 30 Thalassemia patients, including Interview and medical record data from a hospital in Surabaya, Indonesia	Simplifies early detection of Thalassemia Major through easy interpretation and automatic feature selection; effective even with smaller datasets	Relies on binary Yes/No symptom data, which may oversimplify complex medical conditions; potential risk of overfitting despite high accuracy	C4.5 showed the best performance with 100% accuracy

D. Discussion

Machine learning (ML) models have notably enhanced diagnostic accuracy and efficiency. Despite these advancements, there are inherent challenges and areas for future research that are crucial for the evolution and integration of these technologies into clinical practice.

1. Integration of ML in Clinical Workflows

A significant advancement is the integration of ML models into existing clinical workflows, which promises to streamline diagnostic processes and improve patient outcomes. For instance, the use of convolutional neural networks

(CNNs) for analyzing blood smear images has shown high diagnostic accuracy. However, the adoption of ML tools in clinical settings often encounters challenges, including skepticism from healthcare professionals regarding the reliability and transparency of these tools. To foster broader acceptance, future research should focus on enhancing the interpretability of ML models and developing user-friendly interfaces that facilitate their integration into routine clinical practice.

2. Tackling Data Diversity and Model Generalization

This review underscores the critical issue of data diversity in training ML models. Many studies utilize datasets that may not be representative of diverse populations, potentially leading to biased and inaccurate diagnostic outcomes when applied globally. Addressing this, future initiatives must prioritize the collection and analysis of varied datasets that encompass a wider demographic. This approach would help in building more robust models capable of delivering reliable diagnostics across different ethnicities and genetic backgrounds.

3. Advancements in Deep Learning Technologies

Deep learning, particularly through sophisticated image recognition and analysis, offers profound potential for diagnosing Thalassemia from medical imaging. However, these technologies demand substantial computational resources and extensive datasets for training, which can be a barrier in resource-limited settings. Research should thus not only pursue the refinement of these algorithms to improve efficiency and reduce computational demands but also explore innovative training paradigms

4. Ethical and Privacy Considerations

As ML applications become more prevalent in healthcare, ethical and privacy concerns related to the use of sensitive genetic and health data come to the forefront. It is imperative that these technologies are developed and implemented with stringent adherence to ethical standards and privacy regulations to protect patient information. Future frameworks and policies should aim to balance innovation in ML applications with assurances of data security and patient confidentiality

E. Conclusion

This systematic review rigorously examines the intersection of machine learning (ML) technologies and Thalassemia diagnostics, emphasizing the substantial progress and notable achievements in the field. Key outcomes from the application of various ML algorithms indicate a transformative potential in the diagnosis and management of Thalassemia, enhancing both accuracy and efficiency. Algorithms such as AdaBoost and deep learning models have proven effective in detecting intricate disease patterns that traditional methods might miss, achieving high accuracy rates and reducing the invasiveness of diagnostic procedures.

However, despite these technological advances, several challenges remain prevalent. These include the need for larger, more diverse datasets to train the models effectively and ensure their applicability across different demographics. Moreover, the transparency and interpretability of ML models remain critical concerns. The ability of practitioners to understand and trust these models is paramount for their integration into clinical workflows.

Furthermore, while the review covers a broad spectrum of ML applications, it also highlighted the necessity for ongoing research. There is a distinct need for studies that not only refine these technologies but also explore their integration into existing healthcare systems globally. Future research should aim to address the current limitations of dataset diversity, model transparency, and integration difficulties.

In conclusion, the integration of ML in Thalassemia diagnostics holds a promising future. It has the potential to revolutionize the healthcare landscape by providing quicker, more accurate diagnoses and by facilitating a shift towards more personalized medicine. The ongoing advancements in ML are likely to expand its applicability not only in Thalassemia but also in other complex hematological disorders, thus broadening the scope of its benefits in medical science.

F. References

[1] R. Origa, "β-Thalassemia," *Genetics in Medicine*, vol. 19, no. 6, pp. 609–619, Jun. 2017, doi: 10.1038/gim.2016.173.

[2] A. Cao and R. Galanello, "Beta-thalassemia," *Genet Med*, vol. 12, no. 2, pp. 61–76, Feb. 2010, doi: 10.1097/GIM.0B013E3181CD68ED.

[3] M. S. Borah, B. P. Bhuyan, M. S. Pathak, and P. K. Bhattacharya, "Machine learning in predicting hemoglobin variants," *Int J Mach Learn Comput*, vol. 8, no. 2, pp. 140–143, Apr. 2018, doi: 10.18178/IJMLC.2018.8.2.677.

[4] Y. K. Fu *et al.*, "The tvgh-nycu thal-classifier: Development of a machine-learning classifier for differentiating thalassemia and non-thalassemia patients," *Diagnostics*, vol. 11, no. 9, p. 1725, Sep. 2021, doi: 10.3390/DIAGNOSTICS11091725/S1.

[5] K. Ferih *et al.*, "Applications of Artificial Intelligence in Thalassemia: A Comprehensive Review," *Diagnostics 2023, Vol. 13, Page 1551*, vol. 13, no. 9, p. 1551, Apr. 2023, doi: 10.3390/DIAGNOSTICS13091551.

[6] S. Liu and S. Liu, "An Application of Machine Learning to Thalassemia Diagnosis," *Journal of Computer and Communications*, vol. 12, no. 2, pp. 211–230, Feb. 2024, doi: 10.4236/JCC.2024.122013.

[7] R. Das *et al.*, "Performance analysis of machine learning algorithms and screening formulae for β–thalassemia trait screening of Indian antenatal women," *Int J Med Inform*, vol. 167, p. 104866, Nov. 2022, doi: 10.1016/J.IJMEDINF.2022.104866.

[8] B. Çil, H. Ayyıldız, and T. Tuncer, "Discrimination of β-thalassemia and iron deficiency anemia through extreme learning machine and regularized extreme learning machine based decision support system," *Med Hypotheses*, vol. 138, p. 109611, May 2020, doi: 10.1016/J.MEHY.2020.109611.

[9] A. Devanath, S. Akter, P. Karmaker, and A. Sattar, "Thalassemia Prediction using Machine Learning Approaches," *Proceedings - 6th International Conference on Computing Methodologies and Communication, ICCMC 2022*, pp. 1166–1174, 2022, doi: 10.1109/ICCMC53470.2022.9753833.

[10] A. J. Zaylaa, M. Makki, and R. Kassem, "Thalassemia Diagnosis Through Medical Imaging: A New Artificial Intelligence-Based Framework," *2022*

International Conference on Smart Systems and Power Management, IC2SPM 2022, pp. 41–46, 2022, doi: 10.1109/IC2SPM56638.2022.9988891.

[11] B. Nair, C. Mysorekar, R. Srivastava, and S. Kale, "Towards thalassemia detection using optoelectronic measurements assisted with machine-learning algorithms: a non-invasive, pain-free and blood - free approach towards diagnostics," *APSCON 2024 - 2024 IEEE Applied Sensing Conference, Proceedings*, 2024, doi: 10.1109/APSCON60364.2024.10466125.

[12] W. Xu, Y. Song, and T. Zou, "Prediction of Thalassemia Based on SAELM Hybrid Algorithm," *Proceedings - 2019 3rd International Conference on Data Science and Business Analytics, ICDSBA 2019*, pp. 227–230, Oct. 2019, doi: 10.1109/ICDSBA48748.2019.00054.

[13] F. Akhtar, A. Shakeel, J. Li, Y. Pei, and Y. Dang, "Risk Factors Selection for Predicting Thalassemia Patients using Linear Discriminant Analysis," *Proceedings - 2020 Prognostics and Health Management Conference, PHM-Besancon 2020*, pp. 1–7, May 2020, doi: 10.1109/PHM-BESANCON49106.2020.00008.

[14] 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, doi: 10.1109/ACCESS.2021.3066782.

[15] A. R. Laeli, Z. Rustam, S. Hartini, F. Maulidina, and J. E. Aurelia, "Hyperparameter Optimization on Support Vector Machine using Grid Search for Classifying Thalassemia Data," *2020 International Conference on Decision Aid Sciences and Application, DASA 2020*, pp. 817–821, Nov. 2020, doi: 10.1109/DASA51403.2020.9317227.

[16] S. Purwar, R. Tripathi, R. Ranjan, and R. Saxena, "Classification of thalassemia patients using a fusion of deep image and clinical features," *Proceedings of the Confluence 2021: 11th International Conference on Cloud Computing, Data Science and Engineering*, pp. 410–415, Jan. 2021, doi: 10.1109/CONFLUENCE51648.2021.9377054.

[17] N. Tressa, A. V, Suma. C. M, S. K. Singh, and S. J, "Alpha Thalassemia Classifier Using Machine Learning Techniques Based on Genetic Mutations," *2023 Third International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, pp. 118–122, Sep. 2023, doi: 10.1109/ICUIS60567.2023.00028.

[18] E. W. Abdulhay, A. G. Allow, and M. E. Al-Jalouly, "Detection of Sickle Cell, Megaloblastic Anemia, Thalassemia and Malaria through Convolutional Neural Network," *Proceedings of 2021 Global Congress on Electrical Engineering, GC-ElecEng 2021*, pp. 21–25, 2021, doi: 10.1109/GC-ELECENG52322.2021.9788131.

[19] A. H. Meti, B. U. Maheswari, and A. Vijjapu, "Advancing Alpha-Thalassemia Carrier Screening for Better Predictions Using Explainable AI," *4th International Conference on Communication, Computing and Industry 6.0, C216 2023*, 2023, doi: 10.1109/C2I659362.2023.10430520.

[20] M. Saleem, W. Aslam, M. I. U. Lali, H. T. Rauf, and E. A. Nasr, "Predicting Thalassemia Using Feature Selection Techniques: A Comparative Analysis," *Diagnostics 2023, Vol. 13, Page 3441*, vol. 13, no. 22, p. 3441, Nov. 2023, doi: 10.3390/DIAGNOSTICS13223441.

[21] R. K. L. ; Ip *et al.*, "Artificial Intelligence-Assisted Diagnostic Cytology and Genomic Testing for Hematologic Disorders," *Cells 2023, Vol. 12, Page 1755*, vol. 12, no. 13, p. 1755, Jun. 2023, doi: 10.3390/CELLS12131755.

[22] K. Phirom, P. Charoenkwan, W. Shoombuatong, P. Charoenkwan, S. Sirichotiyakul, and T. Tongsong, "DeepThal: A Deep Learning-Based Framework for the Large-Scale Prediction of the α+-Thalassemia Trait Using Red Blood Cell Parameters," *Journal of Clinical Medicine 2022, Vol. 11, Page 6305*, vol. 11, no. 21, p. 6305, Oct. 2022, doi: 10.3390/JCM11216305.

[23] F. Zhang, J. Yang, Y. Wang, M. Cai, J. Ouyang, and J. X. Li, "TT@MHA: A machine learning-based webpage tool for discriminating thalassemia trait from microcytic hypochromic anemia patients," *Clinica Chimica Acta*, vol. 545, p. 117368, May 2023, doi: 10.1016/J.CCA.2023.117368.

[24] H. Ayyıldız and S. Arslan Tuncer, "Determination of the effect of red blood cell parameters in the discrimination of iron deficiency anemia and beta thalassemia via Neighborhood Component Analysis Feature Selection-Based machine learning," *Chemometrics and Intelligent Laboratory Systems*, vol. 196, p. 103886, Jan. 2020, doi: 10.1016/J.CHEMOLAB.2019.103886.

[25] S. Y. Lee *et al.*, "Image Analysis Using Machine Learning for Automated Detection of Hemoglobin H Inclusions in Blood Smears - A Method for Morphologic Detection of Rare Cells," *J Pathol Inform*, vol. 12, no. 1, p. 18, Jan. 2021, doi: 10.4103/JPI.JPI_110_20.

[26] S. Diaz-del-Pino, R. Trelles-Martinez, F. A. González-Fernández, and N. Guil, "Artificial intelligence to assist specialists in the detection of haematological diseases," *Heliyon*, vol. 9, no. 5, p. e15940, May 2023, doi: 10.1016/J.HELIYON.2023.E15940.

[27] P. Feng *et al.*, "An online alpha-thalassemia carrier discrimination model based on random forest and red blood cell parameters for low HbA2 cases," *Clinica Chimica Acta*, vol. 525, pp. 1–5, Jan. 2022, doi: 10.1016/J.CCA.2021.12.003.

[28] D. Basu, R. Sinha, S. Sahu, J. Malla, N. Chakravorty, and P. S. Ghosal, "Identification of severity and passive measurement of oxidative stress biomarkers for β–thalassemia patients: K-means, random forest, XGBoost, decision tree, neural network based novel framework," *Advances in Redox Research*, vol. 5, p. 100034, Jul. 2022, doi: 10.1016/J.ARRES.2022.100034.

[29] D. Mo, Q. Zheng, B. Xiao, and L. Li, "Predicting thalassemia using deep neural network based on red blood cell indices," *Clinica Chimica Acta*, vol. 543, p. 117329, Mar. 2023, doi: 10.1016/J.CCA.2023.117329.

[30] A. Karollus, Ž. Avsec, and J. Gagneur, "Predicting mean ribosome load for 5'UTR of any length using deep learning," *PLoS Comput Biol*, vol. 17, no. 5, May 2021, doi: 10.1371/JOURNAL.PCBI.1008982.

[31] V. Laengsri, W. Shoombuatong, W. Adirojananon, C. Nantasenamart, V. Prachayasittikul, and P. Nuchnoi, "ThalPred: A web-based prediction tool for discriminating thalassemia trait and iron deficiency anemia," *BMC Med Inform Decis Mak*, vol. 19, no. 1, pp. 1–14, Nov. 2019, doi: 10.1186/S12911-019-0929-2/FIGURES/6.

[32] J. Zhang *et al.*, "A MALDI-TOF mass spectrometry-based haemoglobin chain quantification method for rapid screen of thalassaemia," *Ann Med*, vol. 54, no. 1, pp. 293–301, Dec. 2022, doi: 10.1080/07853890.2022.2028002.
[33] D. C. Tran *et al.*, "PREVALENCE OF THALASSEMIA IN THE VIETNAMESE POPULATION AND BUILDING A CLINICAL DECISION SUPPORT SYSTEM FOR PRENATAL SCREENING FOR THALASSEMIA," *Mediterr J Hematol Infect Dis*, vol. 15, no. 1, pp. e2023026–e2023026, Apr. 2023, doi: 10.4084/MJHID.2023.026.
[34] A. Rodríguez-González *et al.*, "A New Artificial Intelligence Approach Using Extreme Learning Machine as the Potentially Effective Model to Predict and Analyze the Diagnosis of Anemia," *Healthcare 2023, Vol. 11, Page 697*, vol. 11, no. 5, p. 697, Feb. 2023, doi: 10.3390/HEALTHCARE11050697.
[35] Y. Zhang, Z. Han, Q. Gao, X. Bai, C. Zhang, and H. Hou, "Prediction of K562 Cells Functional Inhibitors Based on Machine Learning Approaches," *Curr Pharm Des*, vol. 25, no. 40, pp. 4296–4302, Nov. 2019, doi: 10.2174/1381612825666191107092214.
[36] M. Badat *et al.*, "Direct correction of haemoglobin E β-thalassaemia using base editors," *Nat Commun*, vol. 14, no. 1, Dec. 2023, doi: 10.1038/S41467-023-37604-8.
[37] F. Rustam *et al.*, "Prediction of [... formula ...]-Thalassemia carriers using complete blood count features," *Sci Rep*, vol. 12, no. 1, p. 19999, Dec. 2022, doi: 10.1038/S41598-022-22011-8.
[38] S. Uçucu and F. Azik, "Artificial intelligence-driven diagnosis of β-thalassemia minor & iron deficiency anemia using machine learning models," *J Med Biochem*, vol. 43, no. 1, p. 11, Jan. 2024, doi: 10.5937/JOMB0-38779.
[39] A. Sani *et al.*, "Diagnosis and screening of abnormal hemoglobins," *Clinica Chimica Acta*, vol. 552. 2024. doi: 10.1016/j.cca.2023.117685.
[40] M. Ibrahim *et al.*, "Fuzzy-Based Fusion Model for β -Thalassemia Carriers Prediction Using Machine Learning Technique," *Advances in Fuzzy Systems*, vol. 2024, 2024, doi: 10.1155/2024/4468842.
[41] Y. Long and W. Bai, "Constructing a novel clinical indicator model to predicate the occurrence of thalassemia in pregnancy through machine learning algorithm," *Frontiers in Hematology*, vol. 3, p. 1341225, Apr. 2024, doi: 10.3389/FRHEM.2024.1341225.
[42] A. Jahan, G. Singh, R. Gupta, N. Sarin, and S. Singh, "Role of Red Cell Indices in Screening for Beta Thalassemia Trait: an Assessment of the Individual Indices and Application of Machine Learning Algorithm," *Indian Journal of Hematology & Blood Transfusion*, vol. 37, no. 3, p. 453, Jul. 2021, doi: 10.1007/S12288-020-01373-X.
[43] H. Setiawan *et al.*, "Implementation of Fuzzy-based Model for Prediction of Thalassemia Diseases," *J Phys Conf Ser*, vol. 1751, no. 1, p. 012034, Jan. 2021, doi: 10.1088/1742-6596/1751/1/012034.
[44] H. Setiawan *et al.*, "Classification of thalassemia data using random forest algorithm," *J Phys Conf Ser*, vol. 1490, no. 1, p. 012050, Mar. 2020, doi: 10.1088/1742-6596/1490/1/012050.
[45] S. Uçucu, T. Karablylk, and F. M. Azik, "Machine learning models can predict the presence of variants in hemoglobin: Artificial neural network-based

recognition of human hemoglobin variants by HPLC," *Turkish Journal of Biochemistry*, vol. 48, no. 1, pp. 5–11, Feb. 2022, doi: 10.1515/TJB-2022-0093/MACHINEREADABLECITATION/RIS.

[46] A. M. Hussein, A. Sharaf-Eldin, A. Abdo, and S. M. Kamal, "Hybrid Model for Prediction of Treatment Response in Beta-thalassemia Patients with Hepatitis C Infection," *Lecture Notes in Networks and Systems*, vol. 224, pp. 561–584, 2022, doi: 10.1007/978-981-16-2275-5_37.

[47] Y. Setiawan, O. A. Permata, and M. P. Yuda, "Decision Tree based Data Modelling for First Detection of Thalassemia Major," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 13, no. 1, pp. 49–56, Feb. 2024, doi: 10.32736/SISFOKOM.V13I1.1949.