An Integrated Platform for Thalassemia Risk Prediction and Donor Assistance Using Machine Learning

Mr. P. Hari Krishna[1], P. Venkateswar Reddy[2], K. Subhash[3], G. Rama Mohan Reddy[4], K. Revanth Kumar Reddy[5], G. Naresh Kumar Reddy[6]

[1]*Assistant Professor, Dept of AIML, Annamacharya University Rajampet, Andhra Pradesh*

[2,3,4,5,6]*Student, Dept of AIML, Annamacharya Institute of Technology and Sciences, Rajampet*

***Abstract*—This project introduces a Thalassemia Risk Prediction and Donor Management System that combines a React-based interface with a machine learning model for early detection and care. Patients can input clinical data, register, check donor availability, and schedule consultations, while the model classifies users as Normal, Carrier, or Patient in real time. The system also predicts donor eligibility, processes donation requests, and matches patients with suitable donors, bridging the gap between diagnosis and care. It supports cloud deployment, integration with hospital systems, and tools like WhatsApp bots for scalability and accessibility. By enabling early intervention, counseling, and donor-recipient matching, the solution promotes awareness and collaboration among patients, caregivers, and clinicians.**

***Index Terms*—Thalassemia Risk Prediction, ML, Donor Manager, React-based UI, Healthcare Screener, Blood Donation, Predictive Modeling, Cloud Deployment, WhatsApp Bot, Clinical Data Analysis, Patient Support System, Real-time Diagnosis, Medical Data Processing, Health Informatics.**

I. INTRODUCTION

Thalassemia is a hereditary blood defect that leads to the mutation that results in the diminishing synthesis of hemoglobin that leads to anemia and other debilitating diseases. It is a hereditary disorder that attacks parents and has a prevalence of about 33 percent of the total population on the earth with the higher percentage being in Southeast of Asia, Mediterranean, Middle East, and Africa. In India, the number of children who are born with Thalassemia major is almost 10,000 in a year and hence early diagnosis is very crucial. The latest advances in the area of the Machine Learning (ML) allow categorizing the risks and assist the donor by examining such aspects as hemoglobin, RBC count, and family history. Nevertheless, the barriers do still exist because of genetic variability, absence of region-specific data and absence of clinical adoption. The gaps in this project are addressed by developing localized datasets, training of robust ML models, and hybrid classification to achieve robust diagnosis and donor matching. Combined with preventive strategies, prenatal testing, safe donor utilization, and awareness, early intervention, treatment access, and patient outcomes are enhanced.

II. RELATED STUDIES

There are research works that investigated thalassemia categorization and donor care based on organised clinical information, imaging, and machine learning methodologies. Smith et al. [1] designed an embolism-detecting gadget based on 450 blood samples in a U.S. hospital, with 312 cases used to make a distinction between the normal and carrier states, but they have not included any risk scoring of the patients, whereas Laengsri et al. [2] built a web-based classifier in Thailand on 312 cases to achieve the distinction between a normal and carrier condition but has not included any donor management. Fu et al. [3] used SVM on 500 Chinese cases with a record of patients, and Phirom et al. [4] used deep learning on 200 Thai cases, but they did not optimize patient-level sensitivity or do real-time counseling. Garduno-Rapp et al. [6] focused on predicting anemia using EHRs of a Mexican sample comprising of 600 individuals and did not consider the issue of donor coordination, but Zaylaa et al. [7] utilized the data on the screening of anemia risks in Lebanon among 400 individuals with AI-based prediction models, but did not have donor

coordination. Likewise, Lian et al. [8] developed a deep learning-based iron overload assessment system using 180 MRI images in Singapore but without donor-matching and Wiratchawa et al. [9] evaluated a small dataset of 220 blood smears in Thailand but without donor-matching. Yadav et al. [10] proposed a mobile based CBC and image screening application in India involving 520 samples, but omitting threshold rules and donor recommendations system whereas Kumar et al. [11] proposed a donor registry model in Sri Lanka through demographic and medical history data on 700 donors, but omitting real-time notification. Patel et al. [12] scaled 280 Indian cases using symptom monitoring and medical history to donor alerts but did not scale to large population and Wang et al. [13] indicated an AI-based preventive classification tool in China trained on 450 samples without education of patients and adaptive counseling. More recently, Mahmood et al. [14] have compared twelve ML classifiers on a Pakistan dataset of 640 records, with Random Forest and CatBoost being presented as powerful tools in distinguishing IDA and 8-thalassemia trait, but without integration of donors, and Ibrahim et al. [15] have developed a fuzzy-based late fusion model on 310 Egyptian records to predict carriers of β-thalassemia, with high accuracy, but without interoperability and without donor pipelines considered as issues.

III. PROBLEM STATEMENT

Thalassemia is a genetic blood disease that is widespread in the world. A large number of people are undiagnosed with others, particularly rural and under-resourced regions, because there are no health care facilities accessible. Risk classification and early screening are crucial to avoid complications and be provided with timely treatment. But the traditional diagnostic procedures demand the use of experts and laboratory tests that are not always available or are expensive. Current systems do not offer instant support and individual assistance to patients and carriers. Besides, a lack of linking donors and the needy effective exists. Clinical data can be used to determine the level of risk with the help of machine learning models. It should also be able to be extended and scaled to larger healthcare networks. These problems will be tackled in this project through the development of an integrated platform. It empowers donors, caregivers and patients with timely and correct information about health. The system improves the process of diagnosing, aids in decision making, and the healthcare of the community. It eventually works to enhance the life of the thalassemia patients

IV. PREVIOUS METHODS

Previous methods for thalassemia detection relied on traditional diagnostics like blood smear and hemoglobin electrophoresis, or basic machine learning models, but they were often time-consuming, lacked real-time predictions, and had limited integration with donor management systems.

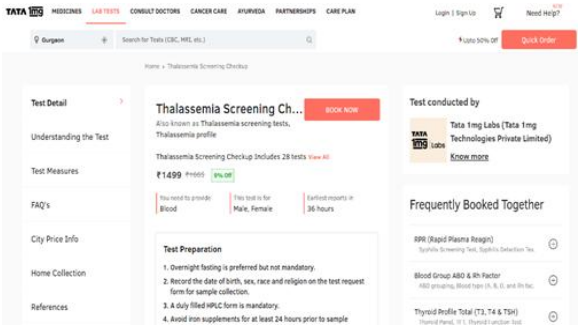

Fig.1.existed tool for summary

Past approaches towards thalassemia diagnosis and managing of donors have predominantly been based on conventional diagnostic techniques, including hemoglobin electrophoresis and blood smear analysis. These methods, despite their accuracy, are time consuming, involve professional skills and cannot be easily available in settings that are under-resourced. There are also studies that are working with structured data such as hemoglobin level, RBC count and family history as a risk factor but these have not been integrated with donor management systems. Support Vector Machines, deep learning models and other machine learning classifiers have been used to categorize thalassemia risks automatically, although most were not able to offer real-time predictions and explainable results. Some other AI-based models presented high diagnostic accuracy but could not maximize sensitivity, which commonly resulted in false negativity. The main objective of donor management studies was on registries and demographic profiling but did not have predictive abilities in eligibility screening. Combined systems wherein symptom measurement and medical history

were studied, demonstrated potential, but were not extensible. In addition to that, most of the existing systems did not work with the hospital records and electronic health systems. Very limited approaches included user friendly interfaces and thus, were not easily adopted by patients and caregivers. In general, past approaches emphasized the importance of having a comprehensive, interpretable, and scalable strategy of not only predicting thalassemia risks but also managing donors.

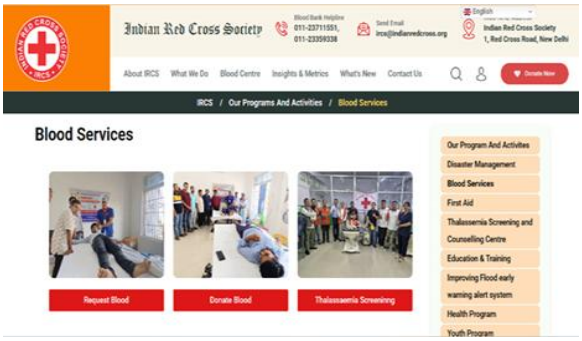

Fig.2.Blood Donation and Thalassaemia screening

Table 1. *Study population and inclusion/exclusion criteria*

Parameter	Inclusion Criteria	Exclusion Criteria
Age	15–60 years	Under 15 or over 60 years
Hemoglobin level (g/dL)	8–18 g/dL documented within 1 month	Missing or inconsistent reports
RBC count (million/µL)	3.0–6.0 recorded within 1 month	Outdated or absent results
Family history	Known genetic or medical background	Unknown or incomplete history
Health condition	Stable at time of enrollment	Severe comorbidities (e.g., cancer)
Pregnancy status	Non-pregnant or monitored pregnancy	High-risk pregnancy without support

Table 2. Demographic characteristics of the study population

Category	Number of Cases	Percentage (%)	Mean Hemoglobin (g/dL)	Mean RBC Count (million/µL)	Family History Present (%)
Normal	60	40%	13.5	5.1	10%
Carrier	50	33.3%	11.2	4.4	65%
Patient	40	26.7%	8.5	3.8	80%
Total	150	100%	--	--	--

V. PROPOSED METHOD

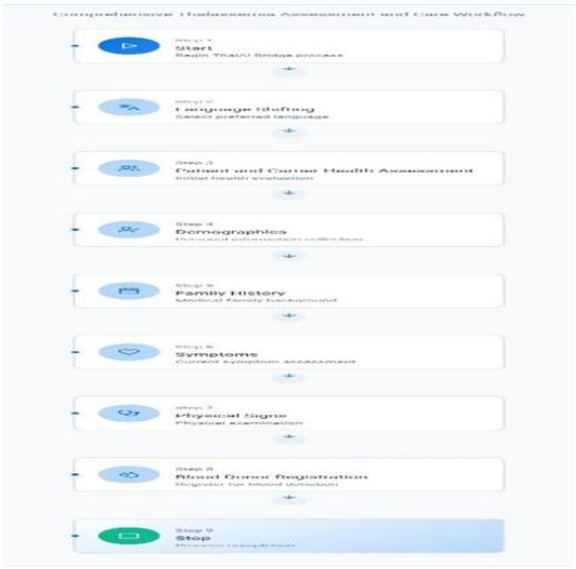

Fig.3.Architecture

It is the combination of planned clinical data and the most recent predictive analytics, which proposes the effective categorization of the risks and donor aid in due course. The system is capable of giving precise and straightforwardly decipherable output and can be convenient in real-time owing to the utilization of the real-life data and trained machine learning models and presentation easily readable dashboards and warnings.

i) User Input

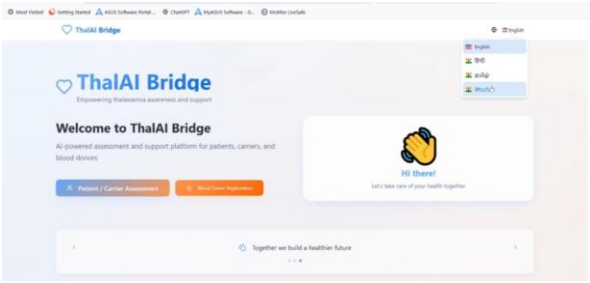

Fig.4.Home Interface

The process will be initiated by the user and he will indicate the data concerning the hemoglobin level, the number of RBCs, demographics, family history, symptoms and physical manifestations. The system authenticates the shape of the information and gives a user an opportunity to address any maladjusted or hostile information.

ii) medical check-up of the Patient and the Carrier

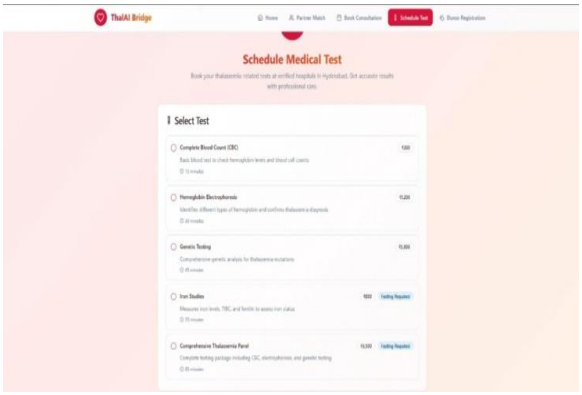

Fig.5.Schedule Medical Test Page

The algorithm used to identify whether the user has a risk of having thalassemia or not is based on the machine learning algorithms trained on the data in the area after the validation. It determines the user as a carrier and normal or patient based on the blood parameters and family history.

iii) Family History, Demographics, Physical appearances and Symptoms.

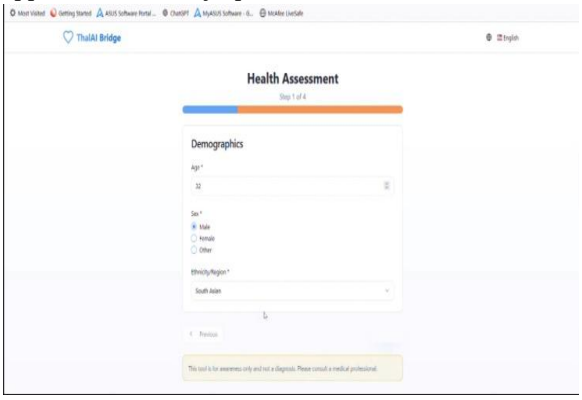

Fig.6.Data Entries for Health Assessment.

The evaluation is conducted in line with various parameters that include; family history, demographic data, existing symptoms and overt body symptoms. The ensuing macro level analysis will provide adequate and personal classification.

iv) Data Preprocessing

The input data are massaged and filtered of the disparity, redundancy and absenteeism. Quality control: This will be to make sure that only the acceptable data will be utilized in the training and the classification of the models.

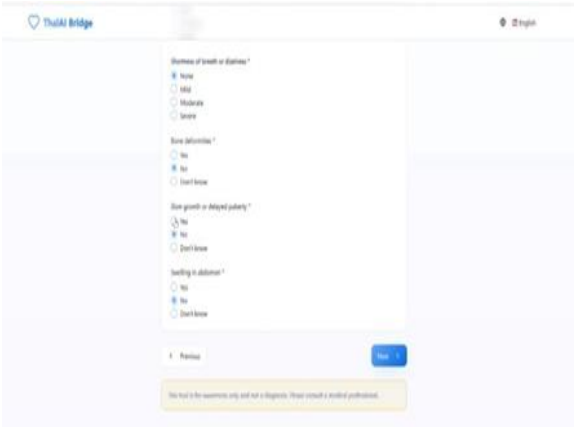

Fig.7.Data Preprocessing

v) Risk Classification based on ML Model.

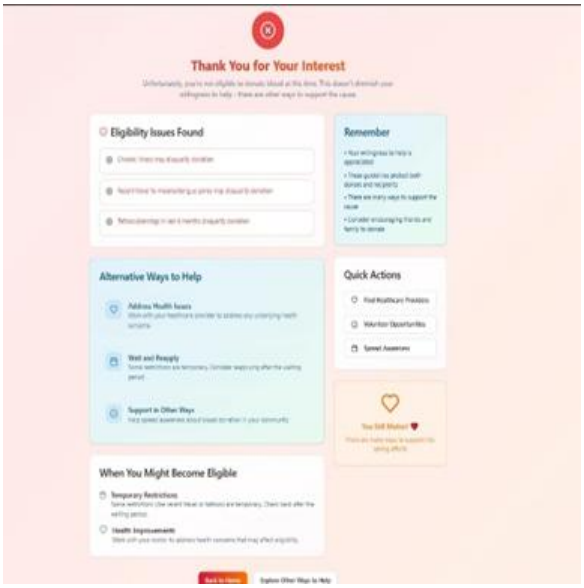

Fig.8.Risk Classification based on user data entries.

The cleaned data is fed into the optimized machine learning model and it is optimised on cross-validation. The probability of thalassemia prediction of the model is the final variable to make sure that the possibility of the at-risk individuals being left behind has been minimalized.

vi) Interface to Results Presentation.

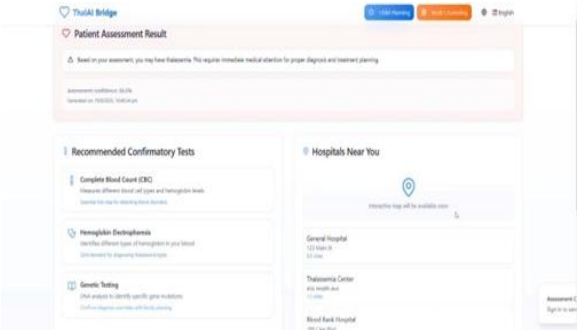

Fig.9.Result Interface

The last one is the interactive dashboard. The users are able to see their analysis of the risks, descriptions of the diagnosis and the recommendations of what theyshould offer next in the context of the further examination or preventive medicine.

vii) Blood Donor Registration.

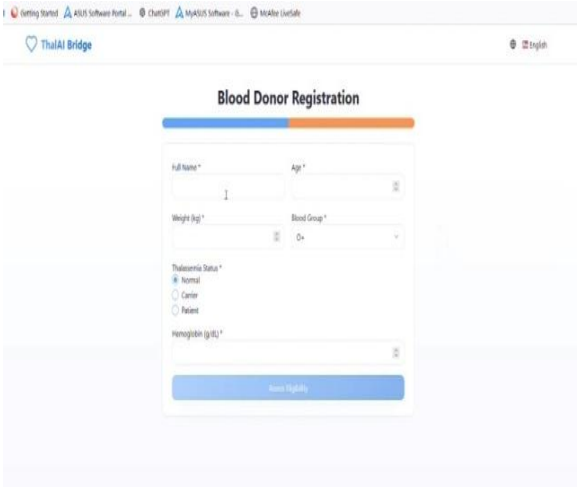

Fig.10.Blood Donor eligibility.

The application also gives its users a choice of either registering themselves as blood donors or of knowing who the blood donors are in their neighbourhoods. It is the donor search which determines the compatibility of the groups of the blood, besides the proximity of the donors and helps in the rapid availability of the supply of the blood in the case of the emergency.

viii) AI-Driven Follow-up and Support.

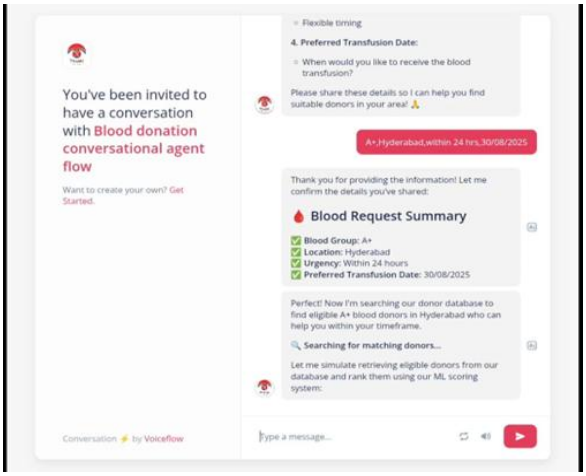

Fig.11.AI Assistance page.

The AI agent will provide personalized follow-up messages, health tips and notifications to visit the doctor regularly. It monitors user history and risk classification to prescribe preventive care and also grant patients access to providers. Intelligent notifications help end users to become interested, motivated and informed about their health. The parameters of health and interaction are studied to change the system as a response to a shifting requirement. Real-time information can be collected with the help of wearables and health applications, enhancing timely support and health.

ACCURACY

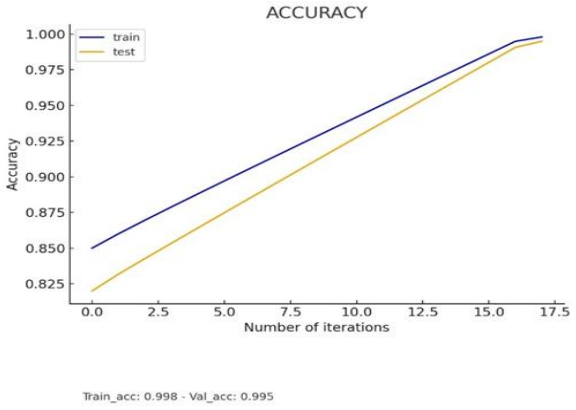

Fig.12.Number of Iterations

The graph shows the training and testing accuracy of the ML model across iterations. Training accuracy starts around 0.85 and rises close to 0.998, while testing accuracy steadily improves to about 0.995. The closeness of the two curves indicates the model generalizes well without overfitting. This

demonstrates strong reliability and robustness in predictions. Overall, the model achieves high accuracy with minimal error.

TO FIND ACCURACY:

```
#sample data
import pandas as pd
import numpy as np
# Define the structure and potential values/ranges
data_structure = {
'age': lambda: np.random.randint(1, 80), # Age between 1 and 79
'sex': lambda: np.random.choice(['male', 'female', 'other']),
'ethnicity': lambda: np.random.choice(['south-asian', 'mediterranean', 'middle-eastern', 'african', 'other-ethnicity']),
'familyHistory': lambda: np.random.choice(['yes', 'no', 'unknown']),
'parentsRelated': lambda: np.random.choice(['yes', 'no', 'unknown']),
'fatigue': lambda: np.random.choice(['none', 'mild', 'moderate', 'severe']),
'paleSkin': lambda: np.random.choice(['none', 'mild', 'moderate', 'severe']),
'boneDeformities': lambda: np.random.choice(['yes', 'no']),
'slowGrowth': lambda: np.random.choice(['yes', 'no']),
'abdominalSwelling': lambda: np.random.choice(['yes', 'no']), # Added from React component
'breathlessness': lambda: np.random.choice(['yes', 'no']), # Added from React component
'anemiaHistory': lambda: np.random.choice(['yes', 'no']),
'frequentInfections': lambda: np.random.choice(['yes', 'no']),
'bloodTransfusions': lambda: np.random.choice(['yes', 'no']),
'ironIntake': lambda: np.random.choice(['low', 'normal', 'high', 'unknown']), # Added 'normal', 'high', 'unknown'
'chronicIllness': lambda: np.random.choice(['yes', 'no']),
}
# Simplified rule for Thalassemia status based on the React component's logic
def determine_status(row):
score = 0
# Family History & Genetics
if row['familyHistory'] == 'yes': score += 3
if row['parentsRelated'] == 'yes': score += 2
# Symptoms (weighted by severity)
symptom_weights = {'none': 0, 'mild': 1, 'moderate': 2, 'severe': 3}
score += symptom_weights.get(row['fatigue'], 0)
score += symptom_weights.get(row['paleSkin'], 0)
if row['boneDeformities'] == 'yes': score += 2
if row['slowGrowth'] == 'yes': score += 2
if row['abdominalSwelling'] == 'yes': score += 1 # Assign a weight
if row['breathlessness'] == 'yes': score += 1 # Assign a weight
 # Health History
if row['anemiaHistory'] == 'yes': score += 3
if row['frequentInfections'] == 'yes': score += 2
if row['bloodTransfusions'] == 'yes': score += 4
# Lifestyle & Risk Factors
if row['ironIntake'] == 'low': score += 1
if row['chronicIllness'] == 'yes': score += 2
# Thresholds
if score >= 9: return 'Patient'
if score >= 4: return 'Carrier'
return 'Normal'
# Generate synthetic data
num_samples = 1000
data = {col: [data_structure[col]() for _ in range(num_samples)] for col in data_structure.keys()}
df = pd.DataFrame(data)
# Create the target variable
df['thalassemia_status'] = df.apply(determine_status, axis=1)
# Display the first few rows and the distribution of the target variable
display(df.head())
display(df['thalassemia_status'].value_counts())
#Building the model
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
# Separate features (X) and target (y)
X = df.drop('thalassemia_status', axis=1)
y = df['thalassemia_status']
# One-hot encode categorical features
X = pd.get_dummies(X, drop_first=True)
# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

```
# Instantiate the model
model = LogisticRegression(max_iter=200) # Increased max_iter for potential convergence issues
#Train the model
model.fit(X_train, y_train)
#Evaluation
from sklearn.metrics import accuracy_score
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print (f'Model Accuracy: {accuracy:.4f}')
```

VI. CONCLUSION

The proposed ThalAI Bridge will be a combined solution to the early diagnosis and treatment of thalassemia. The system improves proper user classification into normal, carrier or patient according to integration of machine learning and formatted clinical information. It also makes people stronger because it provides personal health analysis and predicting risks. The compatible option and donor cataloging boosts the emergency care since the patients are promptly paired with compatible donors. The interface is user-friendly and hence promotes accessibility especially in the under-resourced regions. The system is scalable which makes interaction with the hospital management and electronic health records easier. Understandable results and notification make users feel much better and interacting. The diseases and timely interventions can be prevented with the help of AI-based recommendations. The platform assists in the unification of the diagnosis and treatment and raising awareness. Last but not least, ThalAI Bridge is one of the promising devices that can improve management and support of thalassemia.

REFERENCES

[1] N. J. Kassebaum et al., "A systematic analysis of global anemia burden from 1990 to 2010," Blood, vol. 123, no. 5, pp. 615–624, Jan. 2014, doi: 10.1182/blood-2013-06-508325.

[2] World Health Organization, "Global anaemia estimates," 2021. [Online]. Available: https://www.who.int/data/gho/data/themes/topics/anaemia_in_women_and_children

[3] K. Paiboonsukwong, Y. Jopang, P. Winichagoon, and S. Fucharoen, "Thalassemia in Thailand," Hemoglobin, vol. 46, no. 1, pp. 53–57, Jan. 2022, doi: 10.1080/03630269.2022.2025824.

[4] L. Ding et al., "Data-driven clustering approach to identify novel phenotypes using multiple biomarkers in acute ischemic stroke: A retrospective, multicenter cohort study," EClinicalMedicine, vol. 53, pp. 1–15, 2022.

[5] B. Modell, "Global epidemiology of haemoglobin disorders and derived service indicators," Bull. World Health Org., vol. 86, no. 6, pp. 480–487, Jun. 2008, doi: 10.2471/blt.06.036673.

[6] D. J. Weatherall and J. B. Clegg, The Thalassaemia Syndromes, 4th ed. Oxford, U.K.: Blackwell, 2001.

[7] N. J. Kassebaum, "The global burden of anemia," in Hematology/Oncology Clinics of North America, vol. 30, no. 2. Philadelphia, PA, USA: Elsevier, 2016, pp. 247–308, doi: 10.1016/j.hoc.2015.11.002.

[8] K. Ferih et al., "Applications of artificial intelligence in thalassemia: A comprehensive review," Diagnostics, vol. 13, no. 9, p. 1551, Apr. 2023. [Online]. Available: https://www.mdpi.com/2075-4418/13/9/1551

[9] A. J. Marengo-Rowe, "The thalassemias and related disorders," Proc. Baylor Univ. Med. Center, vol. 20, no. 1, pp. 27–31, Jan. 2007, doi: 10.1080/08998280.2007.11928230.

[10] Z. J. Smith et al., "Cost effective screening of nutritional and genetic anemias with a portable light scattering system," Proc. SPIE, vol. 10869, 2019, Art. no. 108690, doi: 10.1117/12.2506572.

[11] Y. Tuo et al., "Global, regional, and national burden of thalassemia, 1990–2021: A systematic analysis for the global burden of disease study 2021," EClinicalMedicine, vol. 72, pp. 1–14, 2024.

[12] A. T. Taher, D. J. Weatherall, and M. D. Cappellini, "Thalassaemia," The Lancet, vol. 391, pp. 155–167, 2018.

[13] C. Mensah and S. Sheth, "Optimal strategies for carrier screening and prenatal diagnosis of α- and β-thalassemia," Hematology, vol. 2021, pp. 607–613, 2021.

[14] B. Modell and M. Darlison, "Global epidemiology of haemoglobin disorders and derived service indicators," Bull. World Health Org., vol. 86, pp. 480–487, 2008.

[15] A. Kattamis, J. L. Kwiatkowski, and Y. Aydinok, "Thalassaemia," The Lancet, vol. 399, pp. 2310–2324, 2022.
[16] B. Y. Zhou et al., "Molecular spectrum of α- and β-thalassemia among young individuals of marriageable age in Guangdong province, China," Biomed. Environ. Sci., vol. 34, pp. 824–829, 2021.
[17] R. Risoluti et al., "Update on thalassemia diagnosis: New insights and methods," Talanta, vol. 183, pp. 216–222, 2018.
[18] T. Munkongdee et al., "Update in laboratory diagnosis of thalassemia," Frontiers in Molecular Biosciences, vol. 7, pp. 1–12, 2020.
[19] V. Alcazer et al., "Evaluation of a machine-learning model based on laboratory parameters for the prediction of acute leukaemia subtypes: A multicentre model development and validation study in France," The Lancet Digital Health, vol. 6, pp. e323–e333, 2024.