Fuzzy-Based Fusion Model for β-Thalassemia Carriers Prediction Using Machine Learning Technique

Muhammad Ibrahim[1], Sagheer Abbas,[2] Areej Fatima,[3] Taher M. Ghazal,[4,5] Muhammad Saleem,[6] Meshal Alharbi,[7] Fahad Mazaed Alotaibi,[8] Muhammad Adnan Khan,[9,10] Muhammad Waqas[1] and Nouh Elmitwally[11,12]

[1]School of Computer Science, National College of Business Administration & Economics, Lahore, 54000, Pakistan.
[2]Department of Computer Science, Bahria University Lahore Campus, Lahore, 54000, Pakistan
[3]Department of Computer Science, Lahore Garrison University, Lahore, Pakistan
[4]Centre for Cyber Physical Systems, Computer Science Department, Khalifa University
[5]Center for Cyber Security, Faculty of Information Science and Technology, UKM, 43600 Bangi, Selangor, Malaysia
[6]School of Computer Science, Minhaj University Lahore, Pakistan
[7]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Alkharj 11942, Saudi Arabia.
[8]Faculty of Computing and Information Technology in Rabigh (FCITR), King Abdulaziz University, Jeddah, Saudi Arabia
[9]School of Information Technology, Skyline University College, University City Sharjah, Sharjah, 1797, UAE
[10]Riphah School of Computing and Innovation, Faculty of Computing, Riphah International University, Lahore Campus, Lahore 54000, Pakistan
[11]School of Computing and Digital Technology, Birmingham City University, Birmingham B4 7XG, UK
[12]Department of Computer Science, Faculty of Computers and Artificial Intelligence, Cairo University, Giza 12613, Egypt

Correspondence should be addressed to Sagheer Abbas; jamsagheer@gmail.com

Abstract

The abnormality of haemoglobin in the human body is the fundamental cause of thalassemia disease. Thalassemia is considered a common genetic blood condition that has received extensive investigation in medical research globally. Likely, inherited disorders will be passed down to children from their parents. If both parents are beta Thalassemia carriers, 25% of their children will have intermediate or major beta thalassemia, which is fatal. An efficient method of beta thalassemia is prenatal screening after couples have received counselling. Identifying Thalassemia carriers involves a costly, time-consuming, and specialized test using quantifiable blood features. However, cost-effective and speedy screening methods must be developed to address this issue. The demise rate due to thalassemia development is outstandingly high around the globe. The passing rate due to thalassemia development can be reduced by following the proper procedure early; otherwise, it significantly impacts the body. A machine learning-based late fusion model proposes the detection of beta-thalassemia carriers by analyzing red blood cells. This study applied the late fusion technique to employ four machine learning algorithms. For identifying the beta thalassemia carriers, Logistics Regression, Naïve Bayes, Decision Tree, and Neural Network, they have achieved an accuracy of 94.01%, 93.15%, 97.93%, and 98.07%, respectively, by using the features-based dataset. The late fusion-based ML model achieved an overall accuracy of 96% for detecting beta-thalassemia carriers. The proposed late fusion model performs better than previously published approaches regarding efficiency, reliability, and precision.

Keywords: Machine Learning (ML), Logistics Regression (LR), Naïve Bayes (NB), Decision Tree (DT), Neural Network (NN), Fuzzy Logic (FL), Internet of medical things (IoMT), Late Fusion model.

Introduction

Thalassemia comes from the Greek terms 'Thalassa' and 'Haima.' "Thalassa" means "the ocean," and "Haima" means "the blood". Thalassemia is a genetic blood disorder characterized by insufficient production of haemoglobin [1]. Haemoglobin plays a crucial role in the human body by transporting oxygen from the lungs to the rest of the body and returning carbon dioxide to the lungs [2].

Thalassemia is one of the world's most frequent diseases, particularly in the Mediterranean. Many countries are currently dealing with the high and rising incidence of thalassemia, which has become a primary public health concern—a significant source of disability and mortality around the world. Early detection of thalassemia can aid in the reduction of death rates. As a result, healthcare practitioners are responsible for making the right decisions. When distinguishing between ordinary people and patients, complete the following options. Who are carriers of diseases, especially when it comes to genetic disorders such as a condition known as thalassemia [3].

There are two divisions of thalassemia based on two polypeptide chains in haemoglobin. These are known as alpha-thalassemia (α) and beta-thalassemia (β). An abnormality causes alpha thalassemia in the alpha polypeptide gene of haemoglobin, whereas beta-thalassemia is caused due to disturbance in the beta polypeptide gene. The development of any of the alpha or beta-thalassemia in a person's body leads to low or abnormal haemoglobin creation in the body [4]. The red blood cells are affected due to inadequate haemoglobin [5].

The classification of thalassemia consists of three stages: major, intermediate, and minor thalassemia. Thalassemia major is the most crucial stage of the disease in which the patient needs a continuous blood transfusion to survive. Thalassemia intermediate is the middle stage of the condition in which the patient needs blood transfusion occasionally. It is also known as mild or moderate anaemia. The patient with thalassemia minor looks physically fit and healthy. They don't need a blood transfusion but maintain their diet and healthy lifestyle [6].

World Health Organization (WHO) identifies that beta-thalassemia has 5.1% carriers worldwide [7]. Many tests are required to diagnose the difference between Iron deficiency anaemia and beta-thalassemia. These tests include serum iron level, Complete blood count, High-performance liquid chromatography, the binding capacity of Iron, and the calculation of ferritin and HBA2. However, these tests are expensive and unavailable everywhere [8].

In many other research disciplines, machine learning approaches are very efficient in producing results. They make managing and analyzing other fields easier and play a significant role in the health sector. A computer-based system can be developed to identify thalassemia with improved accuracy, better results, and more affordable cost. Various machine learning algorithms have offered effective treatments for various biomedical problems. Many models have been presented to analyze the data of other diseases [32,33] like brain tumours [9], kidney diseases [10], lung disorders [11], and Iron deficiency anaemia by using machine learning techniques [25, 26 and 30], including support vector machine [12], K-nearest neighbour [13], fuzzy logic [14, 29 and 31], Deep extreme machine learning [27] and deep neural network [15, 24 and 28].

Logistic regression models a discrete outcome given an input variable. The most popular logistic regression models a binary result (true/false, yes/no, etc.). When analyzing a classification problem, logistic regression is a helpful analysis tool.

Nave Bayes is a superficial learning algorithm that uses the Bayes rule and assumes attributes are class-dependent. Due to its processing efficiency and other benefits, nave Bayes is commonly used in practice [21].

A tree has numerous analogies in real life and has inspired machine learning, classification, and regression. A decision tree can represent the decision analysis process visually and explicitly.

Feature-based data can be handled very effectively by neural network algorithms. Neural networks are computing systems inspired by human brain neural networks [2, 10].

Although machine learning algorithms are currently helpful for identifying illnesses, earlier research models were less accurate because they mainly concentrated on preprocessing methods, data balancing, and other supervised and semi-supervised learning models. A late fusion technique is needed to fuse the accuracy of many machine learning algorithms while maintaining high sickness detection accuracy. This study proposed a late fusion-based ML model that implements Logistics Regression, Naïve Bayes, Decision Tree, and Neural Network for data analysis. The system will use a featured-based dataset of thalassemia reports.

It highlights the importance of accurately predicting β-thalassemia carriers to enable early intervention and genetic counselling. The limitations of existing prediction models and the need for an improved approach are discussed. The objectives of the paper are clearly stated as follows:

1. To develop a fuzzy-based fusion model that combines multiple machine learning algorithms for β-thalassemia carrier prediction.
2. To evaluate the performance of the proposed model using relevant performance metrics and compare it with existing approaches.
3. To analyze the effectiveness of fuzzy logic in improving the accuracy and reliability of β-thalassemia carrier prediction.

The results of four different machine learning algorithms were combined through fuzzification to decide on beta-thalassemia carrier identification. The outcomes demonstrate that the proposed approach is more precise and effective than existing solutions.

Related Work

The goal of the research is to identify thalassemia sickness early. Hirimutugoda and Wijayarathna [2] implemented a three-layer artificial neural network to detect and differentiate malaria and thalassemia. Both diseases are life-threatening and global health issues. Visual inspection of the images of blood analysis taken with a light microscope is a well-known technique for determining malaria and thalassemia. This technique takes much time and is more consuming and expensive. The model used three and four layers of ANN that merged with methods of image analysis to find the accuracy and effectiveness of the classification for identifying the images related to morphological features of the blood erythrocytes. The study claimed that the three-layered ANN approach generated results with an accuracy of 84.54%.

Ayyıldız and Tuncer [5] performed a decision-based diagnosis to identify and discriminate the Iron deficiency anemia (IDA) and beta-thalassemia (β). They implemented red blood cell indices and two effective techniques of machine learning: support vector machine and k-nearest neighbour. Various parameters of Complete blood count were used to differentiate between IDA and β–Thalassemia. Implementation of RBC indices improved the efficiency of the diagnostic model. But if the number of features increases, the system becomes complicated.

Das et al. [16] employed a decision-based system that used decision trees, ANN, and a Naïve Bayes classifier to discriminate β-thalassemia carriers from ordinary people. The Postgraduate Institute of Medical Education and Research in the Indian city of Chandigarh is where the dataset was gathered. Both ratings were determined to be completely sensitive. The screening score for thalassemia characteristics (BTT) was determined to be 79.25 percent and 91.74 percent, respectively, for the combined score of BTT and HbE. Although the mechanism differentiates two main variants related to haemoglobin, it still requires validation with datasets collected from different countries for implementation and unification.

Egejuru et al. [17] implemented a prediction model for identifying the risk of thalassemia disease. The model used supervised machine learning approaches for analyzing the data collected through questionnaires and medical persons. The Waikato Environment for Knowledge Analysis (WEKA) tool was used for data simulation. Identification variables included demographics (age, marital status, gender, social class, and ethnicity) and clinical variables (spleen enlargement, family history, urine colour, diabetes, and inherited disease status). The dataset consisted of 57% disease carriers and 43% non-carriers. The models implemented in the study are multi-layer perception (MLP) and the Naïve Bayes classifier. The study results show that the MLP is a more effective and reliable mechanism for identifying the risk of thalassemia in patients in Nigeria.

Sadiq et al. [18] constructed an ensemble classifier model using a random forest support vector machine and a Gradient boosting machine to identify patients with thalassemia from the Complete blood count (CBC) test data. The model was implemented on the dataset of CBC reports of 5066 patients collected from the Punjab thalassemia prevention program (PTPP). Input parameters used for this study are red blood cells, Haemoglobin, Hematocrit, Mean cell volume, Mean cell haemoglobin concentration, Mean cell haemoglobin, RBC distribution width, platelet count, and white blood cells. The study achieved an accuracy of 93% in identifying β-thalassemia carriers.

Akhtar et al. [19] implemented a Linear discrimination analysis (LDA) classifier to classify the patients with thalassemia using various parameters of a Complete blood count report. The parameters used in the study are Ferritin, HB, RBC, WBC, HCT, and Platelets. The study also used mathematical formulas to discriminate the patients with thalassemia and iron deficiency anaemia. The accuracy achieved 78% results for females and 75% for males.

The fuzzy-based model was developed to classify thalassemia diseases by Susanto et al. [20]. The haemoglobin, MCV, and MCH levels were obtained following the CBC examination to determine the type of thalassemia. Major, Intermedia, Minor, and Not Thalassemia are four output models. The doctor's perspectives on thalassemia were contrasted with the model prediction results against four datasets. Additional data must be used to understand to further test the model's accuracy.

Jahan et al. [21] investigated the research on red cell indices utilizing machine learning techniques, such as an artificial neural network (ANN), to detect Beta-thalassemia traits (BTT) in pregnant women. The optimal cutoff for each index and the BTT detection test characteristics was determined using a Receiver operating characteristic (ROC) curve analysis. The C4.5 and Naive Bayes (NB) classifiers and a back-propagation type ANN were constructed and tested over 3947 patients using the red cell indices. The study emphasizes that none of the red cell features alone helps detect BTT. However, ANN, with a mixture of all red cell indices, exhibited good sensitivity and specificity for this use. Further neural network development might produce a valuable tool for thalassemia screening in remote areas.

Mohammed and Al-Tuwaijari [22] presented various artificial intelligence-based methods and machine learning techniques for classifying and detecting thalassemia utilizing CBC test variables such as RBC, HGB, MCV, HTC, and HB. This system was developed to identify patients with minor thalassemia alpha and major thalassemia beta. The classification methods are decision tree, Naive Bayes, and support vector machine.

Tyas et al. [23] examined multi-layer perceptron to classify erythrocytes present in thalassemia cases. It combined morphological features with texture and colour features to increase the accuracy of erythrocyte classification. The experimental results of 7108 erythrocytes indicated an accuracy of 98.11% for training and 93.77% for testing based on the combination of features. The system's effectiveness was assessed using images captured at various magnifications and on different scanning platforms. The least number of red cells to image for analysis was determined using Poisson modelling, and the results were validated using image sets. Table 1 is showing the comparative analysis with respect to the accuracy of past works that were about anomaly detection in network security.

Table 1: Previous Work Accuracy and Dataset Status

Author	Method	Dataset	Accuracy
Hirimutugoda and Wijayarathna [2]	Artificial Neural Network	Public	86.54%
Sadiq et al. [18]	Random Forest	Private	91%
Sadiq et al. [18]	Support Vector Machine	Private	90%
Sadiq et al. [18]	Gradient Boosting Machine	Private	91%
Susanto et al. [20]	Fuzzy Inference System	Public	89.26%
Jahan et al. [21]	Artificial Neural Network	Private	85.95%
Tyas et al. [23]	Convolutional Neural Network	Public	93.77%

The Aims and objectives of the paper are:

To highlight the importance of identifying beta-thalassemia carriers and their impact on reducing the mortality rate associated with the disease.

To identify the limitations of current screening methods and propose developing cost-effective and speedy screening methods.

To develop a machine learning-based late fusion model for detecting beta-thalassemia carriers by analyzing red blood cells.

To compare the proposed late fusion model's accuracy, efficiency, reliability, and precision with previously published approaches.

To explore the use of machine learning in medical research to detect beta-thalassemia carriers.

To evaluate the performance of four machine learning algorithms, including Logistics Regression, Naïve Bayes, Decision Tree, and Neural Network.

To use a features-based dataset for the development of the late fusion model.

Proposed β-Thalassemia Prediction Model

The Late fusion model based on machine learning is proposed for predicting β-thalassemia carriers. The system used a features-based dataset of thalassemia reports obtained from the Internet of Medical Things (IoMT) enabled devices. The novel features dataset was collected from the Punjab Thalassemia Prevention Program (PTPP) database. Table 2 presents a complete overview of the features.

Table 2: Dataset structure

Sr.	Features	Datatype	Sr.	Features	Datatype
1.	CS/PS	Integer	5.	MCH	Integer
2.	ETC	Nominal	6.	Hb	Integer
3.	Age	Integer	7.	Hct	Integer
4.	Sex	Integer	8.	MCV	Integer

PTPP is the initiative of the Punjab government of Pakistan to protect the people from thalassemia disease. This platform provides support to thalassemia patients in β-Thalassemia carrier screening. Initially, the dataset is divided into training and testing phases. 70% of records were fixed for training and 30% for testing.

The Proposed model consists of training and validation phases. The proposed model consists of various layers that help to diagnose beta thalassemia disease. These layers are data acquisition, preprocessing, and application. The proposed model's first layer is the data acquisition layer, which collects the dataset from PTPP based on IoMT devices [18]. It consisted of twelve variables and a total number of 5066 instants. Output is classified into two categories. The first is β-Thalassemia Non-Carriers, which contains 3051 records, and the second is β-Thalassemia Carriers, with 2015 patient records. The sex distribution ratio is 53% for males and 47% for females.

This unprocessed data may have some missing or noisy values. Normalization of the data and treatment of missing values is accomplished in the preprocessing layer. The normalizing method is used to handle noisy data. In contrast, missing values are driven by calculating existing values' mean and moving averages.

In the training phase, the third layer of the model is the application layer, which predicts thalassemia sickness using four different machine learning algorithms: Logistics Regression (LR), Naïve Bayes (NB), Decision Tree (DT), and Neural Network (NN).

The LR, NB, DT, and NN results are given to the evaluation phase, which calculates the accuracy. It misrates in the targeted class represented by [0, 1], where 0 is for β-Thalassemia non-carrier, and 1 is for β-Thalassemia Carrier investigated. The data is sent to the cloud if the learning criteria are satisfied. Otherwise, it needs to be retrained, as shown in Figure 1.

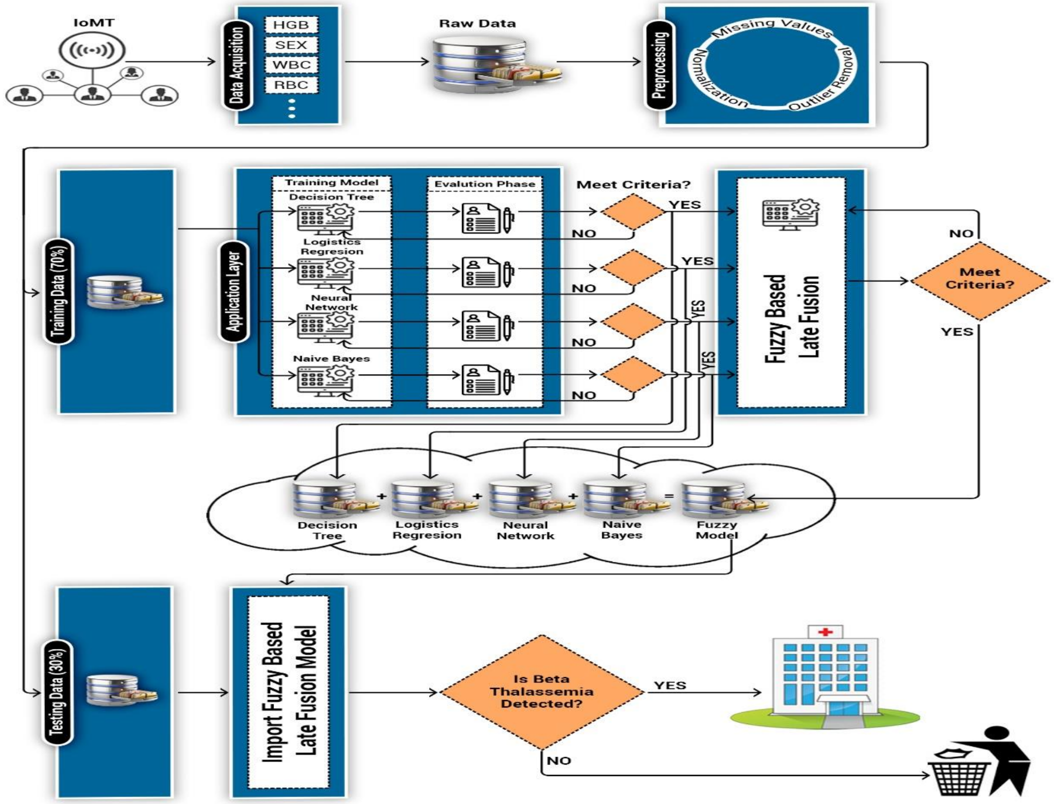


Figure 1: Proposed Late Fusion Model for Thalassemia Disease Prediction

The results of four different techniques are combined in the following stage using a fuzzy inference system to increase the performance of the suggested beta thalassemia carrier's model.

The validation phase utilized the 30% records of the thalassemia dataset to validate the model. The trained fusion-based model is imported from the cloud to predict thalassemia. The model discards the value if a beta-thalassemia non-carrier is found. If a beta-thalassemia carrier is found, the patient is referred to the hospital for additional treatment, as shown in Figure 1.

The following conditions (if-then rules) are employed in the fuzzy logic of the suggested late fusion model, which is written as follows:

The late fusion-based rules identify beta-thalassemia carriers.

$$\mu_{LR\cap \mathrm{NB}\cap DT\cap \mathrm{NN}}(l,n,d,n\ ) = \min[\mu_{LR}(l),\mu_{NB}(n),\mu_{DT}(d),\mu_{NN}(n)]$$

$\boldsymbol{Rule_{bt}^{1}}$= IF (LR is carrier and NB is carrier and DT is carrier and NN is carrier) THEN (Thalassemia is Beta Carrier)

$\boldsymbol{Rule_{bt}^{2}}$= IF (LR is carrier and NB is carrier and DT is carrier and NN is Non-carrier) THEN (Thalassemia is Beta Carrier)

$\boldsymbol{Rule_{bt}^{3}}$= IF (LR is carrier and NB is carrier and DT is Non-carrier and NN is carrier) THEN (Thalassemia is Beta Carrier)

$\boldsymbol{Rule_{bt}^{4}}$= IF (LR is carrier and NB is carrier and DT is Non-carrier and NN is Non-carrier) THEN (Thalassemia is Beta Carrier)

$\boldsymbol{Rule_{bt}^{5}}$= IF (LR is carrier and NB is Non-carrier and DT is carrier and NN is carrier) THEN (Thalassemia is Beta Carrier)

$\boldsymbol{Rule_{bt}^{6}}$= IF (LR is carrier and NB is Non-carrier and DT is carrier and NN is Non-carrier) THEN (Thalassemia is Beta Carrier)

$\boldsymbol{Rule_{bt}^{7}}$= IF (LR is carrier and NB is Non-carrier and DT is Non-carrier and NN is carrier) THEN (Thalassemia is Beta Carrier)

$\boldsymbol{Rule_{bt}^{8}}$= IF (LR is carrier and NB is Non-carrier and DT is Non-carrier and NN is Non-carrier) THEN (Thalassemia is Beta Non-Carrier)

$\boldsymbol{Rule_{bt}^{9}}$= IF (LR is Non-carrier and NB is carrier and DT is carrier and NN is carrier) THEN (Thalassemia is Beta Carrier)

$\boldsymbol{Rule_{bt}^{10}}$= IF (LR is Non-carrier and NB is carrier and DT is carrier and NN is Non-carrier) THEN (Thalassemia is Beta Non-Carrier)

$\boldsymbol{Rule_{bt}^{11}}$= IF (LR is Non-carrier and NB is carrier and DT is Non-carrier and NN is carrier) THEN (Thalassemia is Beta Non-Carrier)

$\boldsymbol{Rule_{bt}^{12}}$= IF (LR is Non-carrier and NB is carrier and DT is Non-carrier and NN is Non-carrier) THEN (Thalassemia is Beta Non-Carrier)

$\boldsymbol{Rule_{bt}^{13}}$= IF (LR is Non-carrier and NB is Non-carrier and DT is carrier and NN is carrier) THEN (Thalassemia is Beta Non-Carrier)

$\boldsymbol{Rule_{bt}^{14}}$= IF (LR is Non-carrier and NB is Non-carrier and DT is carrier and NN is Non-carrier) THEN (Thalassemia is Beta Non-Carrier)

$\boldsymbol{Rule_{bt}^{15}}$= IF (LR is Non-carrier and NB is Non-carrier and DT is Non-carrier and NN is carrier) THEN (Thalassemia is Beta Non-Carrier)

$\boldsymbol{Rule_{bt}^{16}}$= IF (LR is Non-carrier and NB is Non-carrier and DT is Non-carrier and NN is Non-carrier) THEN (Thalassemia is Beta Non-Carrier)

The generated fuzzy rules show that the suggested late fusion-based technique will predict the optimal result based on at least three classification strategies (either the beta-thalassemia carrier or beta-thalassemia non-carrier).

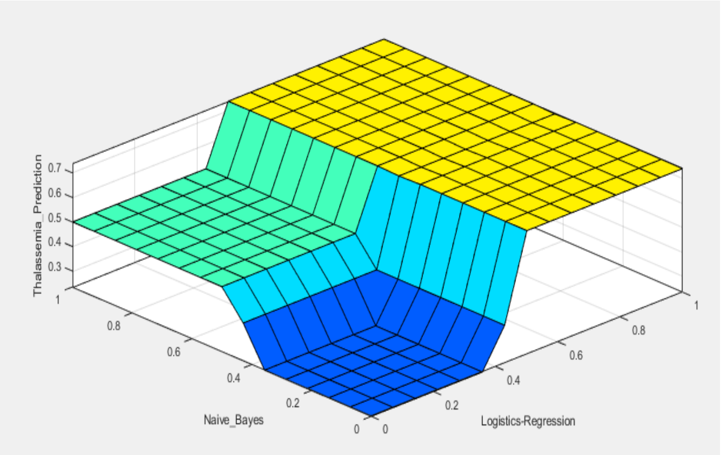


Figure 2: Proposed Late Fusion Rule Surface for NB and LR

The proposed late fusion technique of rule surface for predicting beta-thalassemia carriers based on NB and LR is shown in Figure 2. If both classification methods indicate that "beta thalassemia = carrier" is the outcome, then the suggested technique will also mean that "beta thalassemia = carrier" is the outcome. If both methods indicate that "beta thalassemia = non-carrier" is the outcome, then the proposed technique will suggest that "beta thalassemia = non-carrier" is the outcome.

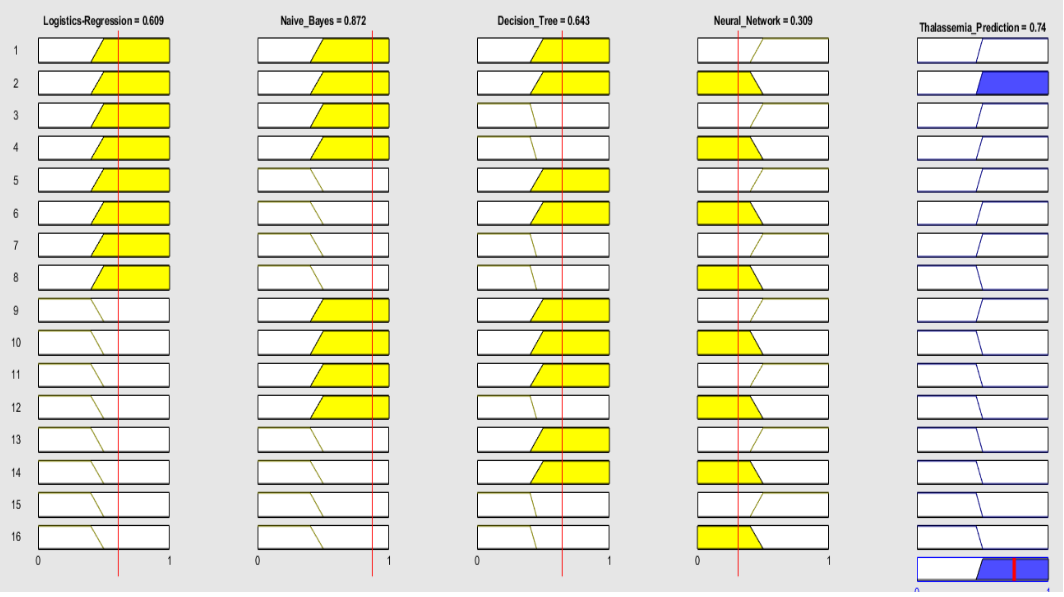


Figure 3: Result of Proposed Late Fusion Model Beta Thalassemia Carrier

Figure 3 demonstrates that the suggested late fusion technique will also predict "beta thalassemia = carrier" if NB, DT, and NN make this prediction 'beta thalassemia = carrier'.

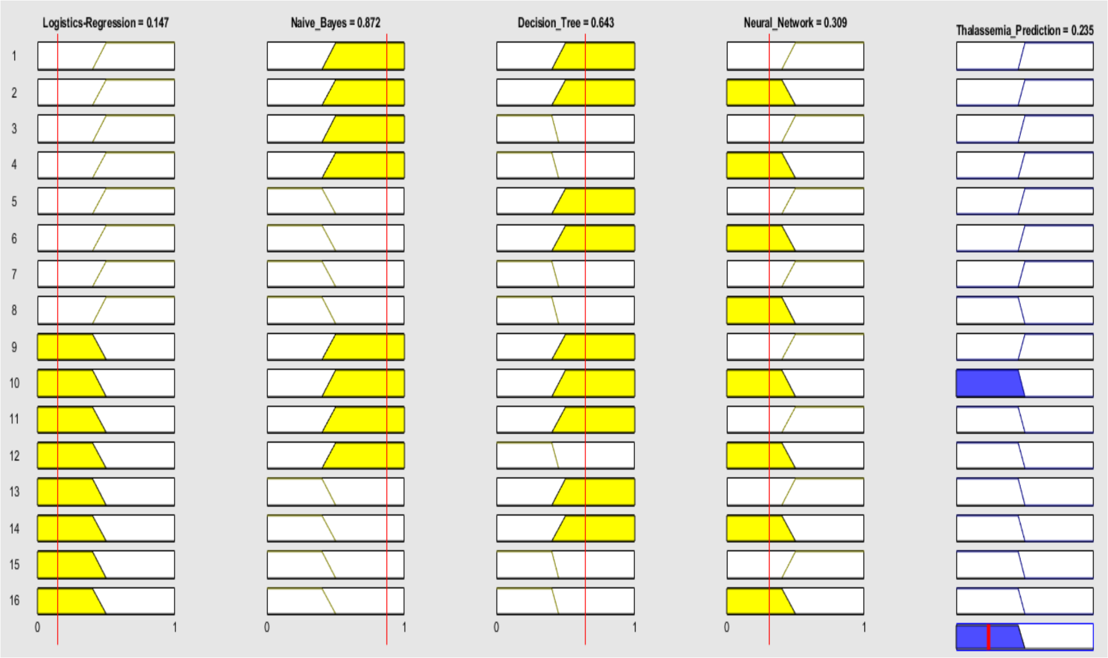


Figure 4: Result of Proposed Late Fusion Model Beta Thalassemia Non-Carrier

Figure 4 shows that if LR and NB show "beta thalassemia = non-carrier," even if DT and NN show "beta thalassemia = carrier," the proposed method will still show "beta thalassemia = non-carrier."

Table 3: Fuzzy-based Graphical and Mathematical Membership Function Representation

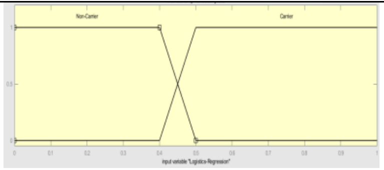
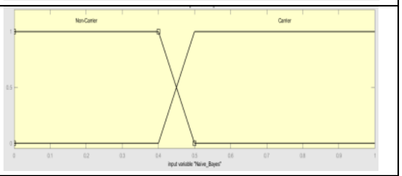
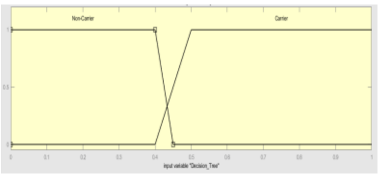
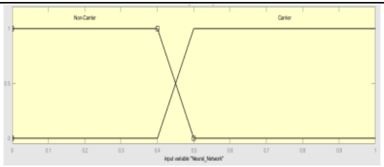
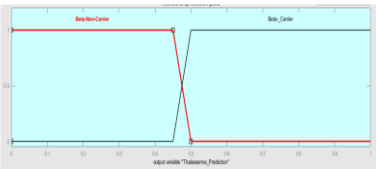
Sr No.	Input / Output Variables	Membership Functions (MF)	Graphical Representation of MF
1	LR = μ_{LR}((lr))	$\mu_{LR,c}(lr) = \left\{max\left(min\left(1,\frac{0.5-\mathrm{lr}}{0.1}\right),0\right)\right\}$ $\mu_{\mathrm{LR,nc}}(lr) = \left\{max\left(min\left(\frac{lr-0.4}{0.1},1\right),0\right)\right\}$	
2	NB = μ_{NB}((nb))	$\mu_{NB,c}(nb) = \left\{max\left(min\left(1,\frac{0.5-\mathrm{nb}}{0.1}\right),0\right)\right\}$ $\mu_{\mathrm{NB,nc}}(nb) = \left\{max\left(min\left(\frac{nb-0.4}{0.1},1\right),0\right)\right\}$	
3	DT= μ_{DT}((dt))	$\mu_{DT,c}(dt) = \left\{max\left(min\left(1,\frac{0.5-\mathrm{dt}}{0.1}\right),0\right)\right\}$ $\mu_{\mathrm{DT,nc}}(dt) = \left\{max\left(min\left(\frac{dt-0.4}{0.1},1\right),0\right)\right\}$	
4	NN = μ_{NN}((nn))	$\mu_{NN,c}(nn) = \left\{max\left(min\left(1,\frac{0.5-\mathrm{nn}}{0.1}\right),0\right)\right\}$ $\mu_{\mathrm{NN,nc}}(nn) = \left\{max\left(min\left(\frac{nn-0.4}{0.1},1\right),0\right)\right\}$	
5	Beta-Thalassemia = μ_{BT}((bt))	$\mu_{BT,c}(bt) = \left\{max\left(min\left(1,\frac{0.5-\mathrm{bt}}{0.05}\right),0\right)\right\}$ $\mu_{\mathrm{BT,nc}}(bt) = \left\{max\left(min\left(\frac{bt-0.45}{0.05},1\right),0\right)\right\}$	

Table 3 shows membership functions based on fuzzy rules. The system testing layer predicts beta-thalassemia carriers. A fuzzy-based cloud model is used to achieve an outcome that stores real-time patient data for evaluation.

Results and Simulation

The late fusion-based model is proposed for the earliest prediction of beta-thalassemia carriers. The results are obtained using MATLAB tool 2022. The proposed model comprises four machine learning techniques, LR, NB, DT, and NN are applied to 5066 features. For both methods, 30% of the fused samples were utilized for validation, while the remaining 70% were used for training. The proposed model diagnoses the beta thalassemia carrier and beta thalassemia non-carrier. The statistical metrics used to evaluate the suggested late fusion model's predicted effectiveness and other categorization methods are explained below. βTc represents beta thalassemia true predicted, βTnc represents beta thalassemia false predicted, βFnc represents beta thalassemia non-carrier false expected, and βFc means expected false beta thalassemia carrier.

$$\text{Accuracy} = \frac{\beta\text{Tc} + \beta\text{Tnc}}{\beta\text{Fc} + \beta\text{Fnc} + \beta\text{Tc} + \beta\text{Tnc}} \tag{1}$$

Accuracy is the number of correctly labelled cases out of the total number of cases.

$$\text{Misrate} = \frac{\beta\text{Fc} + \beta\text{Fnc}}{\beta\text{Fc} + \beta\text{Fnc} + \beta\text{Tc} + \beta\text{Tnc}} \tag{2}$$

The percentage of real positives and negatives missed during an experiment is known as the miss rate.

$$\text{Sensitivity} = \frac{\beta\text{Tc}}{\beta\text{Tc} + \beta\text{Fnc}} \tag{3}$$

Sensitivity measures the capacity of the proposed model to identify positive cases.

$$\text{Specificity} = \frac{\beta\text{Tnc}}{\beta\text{Tnc} + \beta\text{Fc}} \tag{4}$$

$$\text{Positive Predication Value} = \frac{\beta\text{Tc}}{\beta\text{Tc} + \beta\text{Fc}} \tag{5}$$

$$\text{Negative Predication Value} = \frac{\beta\text{Tnc}}{\beta\text{Fnc} + \beta\text{Tnc}} \tag{6}$$

Predictive values, positive and negative, are calculated by dividing each set of results by the proportion of actual successes and failures.

$$\text{False Postive Ratio} = 1 - \frac{\beta\text{Tnc}}{\beta\text{Tnc} + \beta\text{Fc}} \tag{7}$$

$$\text{False Negative Ratio} = 1 - \frac{\beta\text{Tc}}{\beta\text{Tc} + \beta\text{Fnc}} \tag{8}$$

$$\text{Likelihood Ratio Positive} = \frac{\beta\text{Tc}}{\beta\text{Tc} + \beta\text{Fnc}} \Big/ 1 - \frac{\beta\text{Tnc}}{\beta\text{Tnc} + \beta\text{Fc}} \tag{9}$$

$$\text{Likelihood Ratio Negative} = 1 - \frac{\beta\text{Tc}}{\beta\text{Tc} + \beta\text{Fnc}} \Big/ \frac{\beta\text{Tnc}}{\beta\text{Tnc} + \beta\text{Fc}} \tag{10}$$

The dataset contains 5066 instances. 70% of the dataset is used for training which consists of 3,546 records, while the remaining 30% is used for testing, which consists of 1,520 records. The 3546 records were used for training with the LR approach, in which 1715 were beta-thalassemia non-carriers, and 1831 were beta-thalassemia carriers. When trained with LR, 1623 out of 1715 occurrences were non-carriers, while 1717 out of 1831 were found to be carriers. Table 4 displays the results of a comparison between actual and predicted performance throughout training. Results showed an accuracy of 94.2% with a miss rate of 5.8%.

Table 4: Proposed LR-based Training Confusion Matrix

Total Samples = 3546	$O_{BT\text{-}Non\text{-}Carrier}$	$O_{BT\text{-}Carrier}$
$I_{BT\text{-}Non\text{-}Carrier = 1715}$	**1623**	**92**
$I_{BT\text{-}Carrier = 1831}$	**114**	**1717**

In contrast, during the testing of LR, 716 records out of 757 were identified as non-carriers, while 713 records out of 763 were classified as carriers, as shown in Table 5. In LR testing, the attained accuracy was 94.01%, and the miss rate of 5.99%.

Table 5: Proposed LR-based Testing Confusion Matrix

Total Samples = 1520	$O_{BT\text{-}Non\text{-}Carrier}$	$O_{BT\text{-}Carrier}$
$I_{BT\text{-}Non\text{-}Carrier = 757}$	**716**	**41**
$I_{BT\text{-}Carrier = 763}$	**50**	**713**

The 3546 records were used for training with the NB approach, in which 1715 were beta-thalassemia non-carrier, and 1831 were beta-thalassemia carriers. When trained with NB, 1618 out of 1715 occurrences were found to be non-carriers, while 1658 out of 1831 instances were found to be carriers. Table 6 displays the results of a comparison between actual and predicted performance throughout training. Results showed an accuracy of 92.4% with a miss rate of 7.6%.

Table 6: Proposed NB-based Training Confusion Matrix

Total Samples = 3546	$O_{BT\text{-}Non\text{-}Carrier}$	$O_{BT\text{-}Carrier}$
$I_{BT\text{-}Non\text{-}Carrier = 1715}$	**1618**	**97**
$I_{BT\text{-}Carrier = 1831}$	**173**	**1658**

In contrast, during the testing of NB, 721 records out of 757 were identified as non-carriers, while 695 records out of 763 were classified as carriers, as shown in Table 7. In NB testing, the attained accuracy was 93.15% and a miss rate of 6.85%

Table 7: Proposed NB-based Testing Confusion Matrix

Total Samples = 1520	$O_{BT\text{-}Non\text{-}Carrier}$	$O_{BT\text{-}Carrier}$
$I_{BT\text{-}Non\text{-}Carrier = 757}$	**721**	**36**
$I_{BT\text{-}Carrier = 763}$	**68**	**695**

The 3546 records were used for training with the DT approach, in which 1715 were beta-thalassemia non-carrier, and 1831 were beta-thalassemia carriers. When trained with DT, 1703 out of 1715 occurrences were non-carriers, while 1813 out of 1831 were found to be carriers. Table 8 displays the results of a comparison between actual and predicted performance throughout training. Results showed an accuracy of 99.15% with a miss rate of 0.85%.

Table 8: Proposed DT-based Training Confusion Matrix

Total Samples = 3546	$O_{BT\text{-}Non\text{-}Carrier}$	$O_{BT\text{-}Carrier}$
$I_{BT\text{-}Non\text{-}Carrier = 1715}$	**1703**	**12**
$I_{BT\text{-}Carrier = 1831}$	**18**	**1813**

In contrast, during the testing of DT, 756 records out of 757 were identified as non-carriers, while 763 records out of 763 were classified as carriers, as shown in Table 9. In DT testing, the attained accuracy was 99.93% , and a miss rate of 0.07%

Table 9: Proposed DT-based Testing Confusion Matrix

Total Samples = 1520	$O_{BT\text{-}Non\text{-}Carrier}$	$O_{BT\text{-}Carrier}$
$I_{BT\text{-}Non\text{-}Carrier = 757}$	**756**	**1**
$I_{BT\text{-}Carrier = 763}$	**0**	**763**

The 3546 records were used for training with the NN approach, in which 1715 were beta-thalassemia non-carrier, and 1831 were beta-thalassemia carriers. When trained with NN, 1700 out of 1715 occurrences were found to be non-carriers, while 1824 out of 1831 instances were found to be carriers. Table 10 displays the results of a comparison between actual and predicted performance throughout training. Results showed an accuracy of 99.4% with a miss rate of 0.6%.

Table 10: Proposed NN-based Training Confusion Matrix

Total Samples = 3546	$O_{\text{BT-Non-Carrier}}$	$O_{\text{BT-Carrier}}$
$I_{\text{BT-Non-Carrier = 1715}}$	**1700**	**15**
$I_{\text{BT-Carrier = 1831}}$	**7**	**1824**

In contrast, during the testing of NN, 757 records out of 757 were identified as non-carriers, while 763 records out of 763 were classified as carriers, as shown in Table 11. In NN testing, the attained accuracy was 100%.

Table 11: Proposed NN-based Testing Confusion Matrix

Total Samples = 1520	$O_{\text{BT-Non-Carrier}}$	$O_{\text{BT-Carrier}}$
$I_{\text{BT-Non-Carrier = 757}}$	**757**	**0**
$I_{\text{BT-Carrier = 763}}$	**0**	**763**

Table 12 displays detailed results for validation of all used classification machine learning techniques (LR, NB, DT, and NN). It can be observed that all four machine learning techniques performed well and achieved an average accuracy is 96.77% and misrate of 3.23%.

Table 12: ML-based Proposed Model Performance (Validation)

Samples for validation (30% Records)		
Approaches	**Accuracy**	**Miss Rate**
Logistics Regression (LR)	94.01%	5.99%
Naïve Bayes (NB)	93.15%	6.85%
Decision Tree (DT)	99.93%	0.07%
Neural Network (NN)	100%	0%
Average Performance Proposed Model	96.77%	3.23%

Four machine learning techniques are finally provided to the fuzzy system as input for the final prediction. Input to the fuzzy system consists of LR, NB, DT, and NN classifiers and the output class Beta Thalassemia Carriers classifiers. By employing fuzzy rules, the suggested machine learning late fusion-based fuzzy system attained an accuracy of 96% and a miss rate of 4%. The fuzzy system randomly takes twenty-five input ranges for generating the fusion-based results. Based on the fuzzy rules, 12 outputs show beta-thalassemia carriers, and 12 outcomes non-carriers truly predicted. The remaining one is between the carrier and non-carrier stages that, showed the system's error.

Table 13: Comparison of Previous Approaches with Proposed Late Fusion-based ML Model

Author	**Method**	**Accuracy**	**Misrate**
Sadiq et al. [18]	Random Forest	91%	9%
Sadiq et al. [18]	Support Vector Machine	90%	10%
Sadiq et al. [18]	Gradient Boosting Machine	91%	9%
Susanto et al. [20]	Fuzzy Inference System	89.26%	10.74%
Jahan et al. [21]	Naïve Bayes	82.49%	17.51%
Tyas et al. [23]	Convolutional Neural Network	93.77%	6.23%
Proposed Model	**Late fusion-based ML Model**	**96%**	**4%**

Table 13 displays the results of a comparison between the suggested fused machine learning model and the various thalassemia illness prediction methods described in the literature. The proposed late fusion model is compared with RF [18], SVM [18], GBM [18], FIS [20], NB [21], and CNN [23]. Advanced methods are contrasted with the proposed late fusion model. In comparison to the other methods, the proposed late fusion model excelled. The proposed fused model outperformed the different approaches. The suggested machine learning fusion-based system can be included in intelligent healthcare systems for early and accurate beta thalassemia carrier prediction. The proposed model has shown the accuracy of beta thalassemia carrier prediction is 96%.

Conclusions

The critical point of this study is to develop a system to analyze beta-thalassemia carrier patients using the late fusion-based ML model. This system is fundamental and more accessible for medical experts and non-experts. Hence, any person can examine the status of thalassemia just by feeding the required input data. The goal of this study is to analyze the various dimensions of thalassemia. The total precision of this proposed late fusion-based ML model is 96%. The presented framework can be enhanced in the future by utilizing different methods, including federated learning. The study can also be extended by applying Short-Term Long Memory (LSTM) and other ML algorithms and diagnosing the other stages of Thalassemia like Alpha Max and Min, Beta Max, and Min.

Data Availability

Data will provide on demand.

Conflicts of Interest

The authors declare no conflicts of interest to report regarding the present study.

Funding Statement

The authors received no specific funding for this study.

Acknowledgements

Thanks to Gapico PVT, who provide us simulation platform.

References

1. Hossain, M. S., Hasan, M. M., Petrou, M., Telfer, P., & Al Mosabbir, A. (2021). The parental perspective of thalassaemia in Bangladesh: lack of knowledge, regret, and barriers. Orphanet Journal of Rare Diseases, 16(1), 1-10.
2. Hirimutugoda, Y. M., & Wijayarathna, G. (2010). Image analysis system for detection of red cell disorders using artificial neural networks. Sri Lanka Journal of Bio-Medical Informatics, 1(1).
3. Q. Zhuang et al., "The value of combined detection of HbA2 and HbF for the screening of thalassemia among individuals of childbearing ages," Zhonghua yi xue yi Chuan xue za zhi= Zhonghua Yixue Yichuanxue Zazhi= Chinese J. Med. Genet., vol. 39, no. 1, pp. 16–20, 2022.
4. Z. Rustam, A. Kamalia, R. Hidayat, F. Subroto, and A. Suryansyah, "Comparison of Fuzzy C-Means, Fuzzy Kernel C-Means, and Fuzzy Kernel Robust C-Means to Classify Thalassemia Data," Update, vol. 1, p. 1, 2019.

5. Ayyıldız, H., & Tuncer, S. A. (2020). 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, 196, 103886.
6. Nabi, A. T., Muttu, J., Chhaparwal, A., Mukhopadhyay, A., Pattnaik, S. J., & Choudhary, P. (2022). Implications of β-thalassemia on oral health status in patients: A cross-sectional study. Journal of Family Medicine and Primary Care, 11(3), 1174.
7. Sari, D. P., Wahidiyat, P. A., Setianingsih, I., Timan, I. S., Gatot, D., & Kekalih, A. (2022). Hematological Parameters in Individuals with Beta Thalassemia Trait in South Sumatra, Indonesia. Anemia, 2022.
8. Anari, Shokofeh, Nazanin Tataei Sarshar, Negin Mahjoori, Shadi Dorosti, and Amirali Rezaie. "Review of deep learning approaches for thyroid cancer diagnosis." Mathematical Problems in Engineering 2022 (2022).
9. Ranjbarzadeh, R., Bagherian Kasgari, A., Jafarzadeh Ghoushchi, S., Anari, S., Naseri, M., & Bendechache, M. (2021). Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images. Scientific Reports, 11(1), 1-17.
10. V. A. Binson, M. Subramoniam, Y. Sunny, and L. Mathew, "Prediction of pulmonary diseases with electronic nose using SVM and XGBoost," IEEE Sens. J., vol. 21, no. 18, pp. 20886–20895, 2021.
11. Asif, M., Abbas, S., Khan, M. A., Fatima, A., Khan, M. A., & Lee, S. W. (2021). MapReduce based intelligent model for intrusion detection using machine learning technique. Journal of King Saud University-Computer and Information Sciences.
12. Islam, M. M., Rahman, M., Roy, D. C., Islam, M., Tawabunnahar, M., Ahmed, N. A. M., & Maniruzzaman, M. (2022). Risk factors identification and prediction of anemia among women in Bangladesh using machine learning techniques. Current Women's Health Reviews, 18(1), 118-133.
13. Haseli, G., Ranjbarzadeh, R., Hajiaghaei-Keshteli, M., Ghoushchi, S. J., Hasani, A., Deveci, M., & Ding, W. (2023). HECON: Weight assessment of the product loyalty criteria considering the customer decision's halo effect using the convolutional neural networks. Information Sciences, 623, 184-205..
14. Kollias, D., Tagaris, A., Stafylopatis, A., Kollias, S., & Tagaris, G. (2018). Deep neural architectures for prediction in healthcare. Complex & Intelligent Systems, 4(2), 119-131.
15. Das, R., Datta, S., Kaviraj, A., Sanyal, S. N., Nielsen, P., Nielsen, I., ... & Saha, S. (2020). A decision support scheme for beta thalassemia and HbE carrier screening. Journal of advanced research, 24, 183-190.
16. Egejuru, N. C., Olusanya, S. O., Asinobi, A. O., Adeyemi, O. J., Adebayo, V. O., & Idowu, P. A. (2019). Using data mining algorithms for thalassemia risk prediction. J Biomed Sci Eng, 7(2), 33-44.
17. Sadiq, S., Khalid, M. U., Ullah, S., Aslam, W., Mehmood, A., Choi, G. S., & On, B. W. (2021). Classification of β-Thalassemia Carriers From Red Blood Cell Indices Using Ensemble Classifier. IEEE access, 9, 45528-45538.
18. Akhtar, F., Shakeel, A., Li, J., Pei, Y., & Dang, Y. (2020, May). Risk Factors Selection for Predicting Thalassemia Patients using Linear Discriminant Analysis. In 2020 Prognostics and Health Management Conference (PHM-Besançon) (pp. 1-7). IEEE.
19. Susanto, E. R., Syarif, A., Muludi, K., Perdani, R. R. W., & Wantoro, A. (2021). Implementation of Fuzzy-based Model for Prediction of Thalassemia Diseases. In Journal of Physics: Conference Series (Vol. 1751, No. 1, p. 012034). IOP Publishing.
20. Jahan, A., Singh, G., Gupta, R., Sarin, N., & Singh, S. (2021). 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 and Blood Transfusion, 37(3), 453-457.
21. Mohammed, M. Q., & Al-Tuwaijari, J. M. (2021). A Survey on various Machine Learning Approaches for thalassemia detection and classification. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(13), 7866-7871.
22. Tyas, D. A., Hartati, S., Harjoko, A., & Ratnaningsih, T. (2020). Morphological, Texture, and Color Feature Analysis for Erythrocyte Classification in Thalassemia Cases. IEEE Access, 8, 69849-69860.
23. Ihnaini, B., Khan, M. A., Khan, T. A., Abbas, S., Daoud, M. S., Ahmad, M., & Khan, M. A. (2021). A smart healthcare recommendation system for multidisciplinary diabetes patients with data fusion based on deep ensemble learning: computational Intelligence and Neuroscience, 2021.
24. Siddiqui, S. Y., Athar, A., Khan, M. A., Abbas, S., Saeed, Y., Khan, M. F., & Hussain, M. (2020). Modelling, simulation and optimization of diagnosis cardiovascular disease using computational intelligence approaches. Journal of Medical Imaging and Health Informatics, 10(5), 1005-1022.
25. Khan, M. A., Abbas, S., Atta, A., Ditta, A., Alquhayz, H., Khan, M. F., & Naqvi, R. A. (2020). Intelligent cloud based heart disease prediction system empowered with supervised machine learning.
26. Rehman, A., Athar, A., Khan, M. A., Abbas, S., Fatima, A., & Saeed, A. (2020). Modelling, simulation, and optimization of diabetes type II prediction using deep extreme learning machine. Journal of Ambient Intelligence and Smart Environments, 12(2), 125-138.

27. Ahmad, G., Alanazi, S., Alruwaili, M., Ahmad, F., Khan, M. A., Abbas, S., & Tabassum, N. (2021). Intelligent ammunition detection and classification system using convolutional neural network.
28. Fatima, A., Adnan Khan, M., Abbas, S., Waqas, M., Anum, L., & Asif, M. (2019). Evaluation of planet factors of smart city through multi-layer fuzzy logic (MFL). The ISC International Journal of Information Security, 11(3), 51-58.
29. Muhammad, M. U. U. A. H., & Saleem, A. M. S. F. M. Intelligent Intrusion Detection System for Apache Web Server Empowered with Machine Learning Approaches.
30. Naeem, Z., & Naeem, F. (2022). Predicting the performance of governance factor using fuzzy inference system. International Journal of Computational and Innovative Sciences, 1(2), 35-50.
31. Muneer, S., & Rasool, M. A. (2022). AA systematic review: Explainable Artificial Intelligence (XAI) based disease prediction. International Journal of Advanced Sciences and Computing, 1(1), 1-6.
32. Appiahene, P., Asare, J. W., Donkoh, E. T., Dimauro, G., & Maglietta, R. (2023). Detection of iron deficiency anemia by medical images: a comparative study of machine learning algorithms. BioData Mining, 16(1), 1-20.
33. Appiahene, P., Arthur, E. J., Korankye, S., Afrifa, S., Asare, J. W., & Donkoh, E. T. (2023). Detection of anemia using conjunctiva images: A smartphone application approach. Medicine in Novel Technology and Devices, 18, 100237.