Iraqi Journal of Statistical Sciences

http://stats.uomosul.edu.iq

Comparison between Logistic Regression and K-Nearest Neighbour Techniques with Application on Thalassemia Patients in Mosul

Mohammed Faris Ali[1], Hutheyfa H. Taha[2]

[1]Postgraduate Student Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Iraq,[2]Department of Operations Research and Intelligent Technologies, College of Computer Science and Mathematics, University of Mosul, Iraq.

Article information

Article history:
Received: August 15, 2024
Revised: April 1, 2025
Accepted: April 15, 2025
Available: June 1,2025

Keywords:
Machine Learning ,Thalassemia Disease Prediction ,KNN Model

Correspondence:
Mohammed. F. Ali
mohammed.22csp63@student.uomosul.edu.iq

Abstract

Thalassemia is a genetic disease that is transmitted from parents to children when both parents are carriers of the genetic mutation. This change leads to a decrease in the number, quality, and condition of red blood platelets and an increase in the rate of red blood platelet damage, which leads to iron accumulation in the body and a decrease in hemoglobin in the blood. This project aims to develop a model to predict thalassemia using the nearest neighbor technique and the logistic regression model based on the model evaluation criteria: accuracy, recall, precision, F1-score, and AUC. The data were obtained from Al-Hadbaa Specialized Hospital in Mosul. The data set included 280 observations, of which 149 (53.21%) were thalassemia intermedia and 131 (46.78%) were thalassemia major. The data was divided into 70% for training and 30% for screening. The experimental results showed that the logistic regression model performed better than the nearest neighbor algorithm with a precision of 96%, recall of 98%, and F1- score of 97% in the thalassemia intermedia category, while it had a precision of 97%, recall of 95%, and F1- score of 96% in the thalassemia major category, indicating that logistic regression performed well in distinguishing between these two categories. it has been shown that logistic regression is more effective than the K-nearest neighbor algorithm in classifying thalassemia patients, especially those with thalassemia major. The study showed that the type of distance used in the K-nearest neighbor algorithm, whether "Manhattan" or "Chebyshev", has a significant impact on the accuracy of predictions, with the highest accuracy reaching 95% when K = 4. It was also shown that the difference between distance calculation methods and the K value plays a major role in improving the classification results, as it was determined that the optimal value for K is 4, which led to improving the accuracy of predictions. The researcher suggests increasing the data size, as it is possible to improve the accuracy of models by increasing the data size. In addition, the researcher recommends using other artificial intelligence techniques, especially neural networks, to verify any additional improvements.

DOI 10.33899/iqjoss.2025.187789

1. Introduction

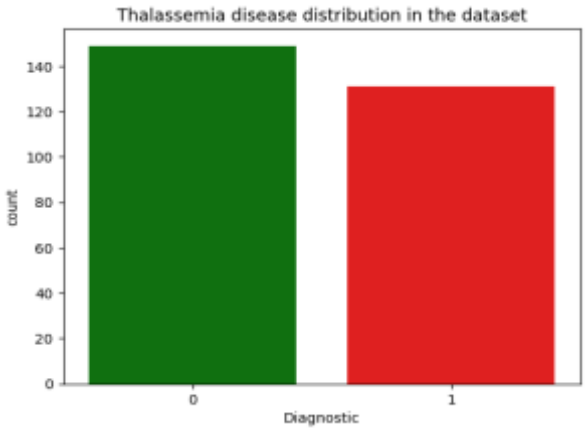
Thalassemia is one of the most common genetic disorders worldwide, especially in the Mediterranean, Middle East, and Southeast Asia. It results from a deficiency in the production of hemoglobin, an essential protein that transports oxygen throughout the bloodstream. As a result, thalassemia patients suffer from chronic anemia and a variety of other health problems. Errors during the diagnostic process are one of the most significant problems in diagnosing thalassemia diseases. These errors occur due to doctors' lack of expertise in effectively detecting thalassemia diseases. In the world of medicine and diagnosis, the precise analysis required to distinguish between the different forms and severity levels of thalassemia is of the utmost importance. The importance of classification lies in its ability to predict outcomes from new data samples and identify

categories that were not part of the training process. The application of artificial intelligence methods to classify thalassemia has greatly increased the accuracy and speed of diagnosis. This research explores the use of supervised learning methods to classify thalassemia patients, which requires the addition of a target variable. In this study, we present a model for predicting thalassemia treatment using the nearest - neighbor approach. Several studies have been undertaken to detect thalassemia illnesses using machine learning techniques. The researchers used several machine learning methods to create a prediction model for identifying thalassemia illnesses. This section discusses some prior studies on the prediction of thalassemia sickness. (Yousefian et al., 2017)) presented a study to predict diabetes in thalassemia patients using the Zafar dataset, with a sample size of 256 observations. The researchers employed KNN and RBFN methods. The findings showed that RBFN performed better than KNN, with an accuracy rate of 81%. (Hartini & Rustam, 2019)) The study presented a new kernel-based technique, derived from hierarchical density-based clustering (HCDP), using the k-nearest neighbor and hierarchical clustering. The study was conducted on data from Harrapan Kita Hospital in Indonesia, which included 82 thalassemia patients and 68 non-thalassemia patients. The results showed that HCDP achieved an F1 score of 67.77%, Which increased to 70.06% when combined with the kernel function (Borah et al., 2018) proposed a study that emphasizes the importance of accurate diagnosis of hemoglobin and thalassemia diseases for effective treatment. They used machine learning techniques such as logistic regression, support vector classifier, nearest neighbor algorithm, naive Bayes algorithm, perceptron classifier, linear support vector classifier, stochastic gradient regression, decision tree, random forest, and multilayer perceptron. They analyzed around 1500 blood samples for hemoglobin classification. KNN, decision tree, and random forest performed the best, with a precision of 93.89%, recall of 92.78%, and F1 score 93.33%, making them the most effective in classifying hemoglobin variants. (Paokanta et al., 2010) conducted a study to compare the performance of different classification approaches by varying the data type to determine the most suitable data type about each approach. We used fJ-thalassemia data to identify the genotype of fJ-thalassemia patients. The results of the study indicate that the data types can serve as a nominal metric for Bayesian networks (BNs) and multinomial logistic regression, achieving accuracy rates of 85.83 and 84.25, respectively. Moreover, the interval metrics can be used well for K-nearest neighbors (KNN), the multilayer neural network (MLP), and naive Bayes with an accuracy of 88.98, 87.40, and 84.25, respectively.

This study used two supervised machine learning algorithms to predict and diagnose beta-thalassemia patients in two groups (thalassemia major and thalassemia intermedia) to select the best classifier based on several model performance evaluation criteria.

2. Research Method

In this research, the researcher collected data on thalassemia patients from Al-Hadbaa Specialized Hospital for Blood Diseases and Bone Marrow Transplantation in Mosul to train and test the KNN model and the logistic regression model. For implementation and experimental testing, the researcher used the Python programming language. In addition, a statistical method, which is Pearson correlation analysis, was used to visualize the data in addition to knowing the relationship between the variables of thalassemia patient's data and the extent of their impact on the dependent variable. Figure (1) shows the distribution of thalassemia disease in the data set. Where the green color represents the number of cases infected with thalassemia intermedia, while the red color represents the number of cases infected with thalassemia major.



.

Figure 1 Number of cases of thalassemia major and intermediate

2.1 Dataset Description

In this study, the thalassemia patient dataset consists of 280 observations and 13 variables. Among the 280 observations, 149 (53.21%) had thalassemia intermedia, whereas 131 (46.78%) had thalassemia major. The dataset has no missing variable values. Table 1 summarizes the thalassemia patient dataset variables. We use 70% of the dataset's observations for training and 30% for testing. Table 1 displays the variables from the thalassemia patient dataset used to train and test the KNN model.

Table 1. Recoding of thalassemia patient data variables

NO.	Variable	Description
1	Gender	Gender (1= Males, 0= Females)
2	Age	The age of a person
3	Splenic enlargement	c enlargement (0= Normal, 1= enlargement, 2= Splenect
4	Heart disease	Heart disease (0= NO, 1= Yes)
5	The growth is delayed	The growth is delayed (0= NO, 1= Yes)
6	Osteoporosis	Osteoporosis (0= NO, 1= Yes)
7	Blood transfusion	Blood transfusion (0= NO, 1= Yes)
8	HB	Hemoglobin blood for Males; HB < 13 Hemoglobin blood for Females; HB <11.5
9	MCV	Mean cell volume; MCV< 80
10	HBA1	Intermediate 50% to 70% Major 0%
11	HBA2	Intermediate 3% to 8% Major 3% to 8%
12	HBF	Intermediate 20% to 40% Major> 90%
13	Diagnostic	Diagnostic; (1=Thalassemia major, 0=Thalassemia Intermediate)

We recorded the descriptive data and converted it to numeric data to standardize the data values by automatic label encoding, which allows the model to easily handle numeric data. Table 1 shows the most important variables that were converted from descriptive to numeric format.

2.2 Correlation Matrix Between Variables Using Heat Maps

Heat maps depict the significance of each variable in the data under investigation, with colors indicating the intensity and strength of each data point in the correlation matrix. The researcher performed a Pearson correlation analysis to determine the link between each of the variables under consideration. This assists in discovering the characteristics that are closely associated with the target group in the data of thalassemia patients.

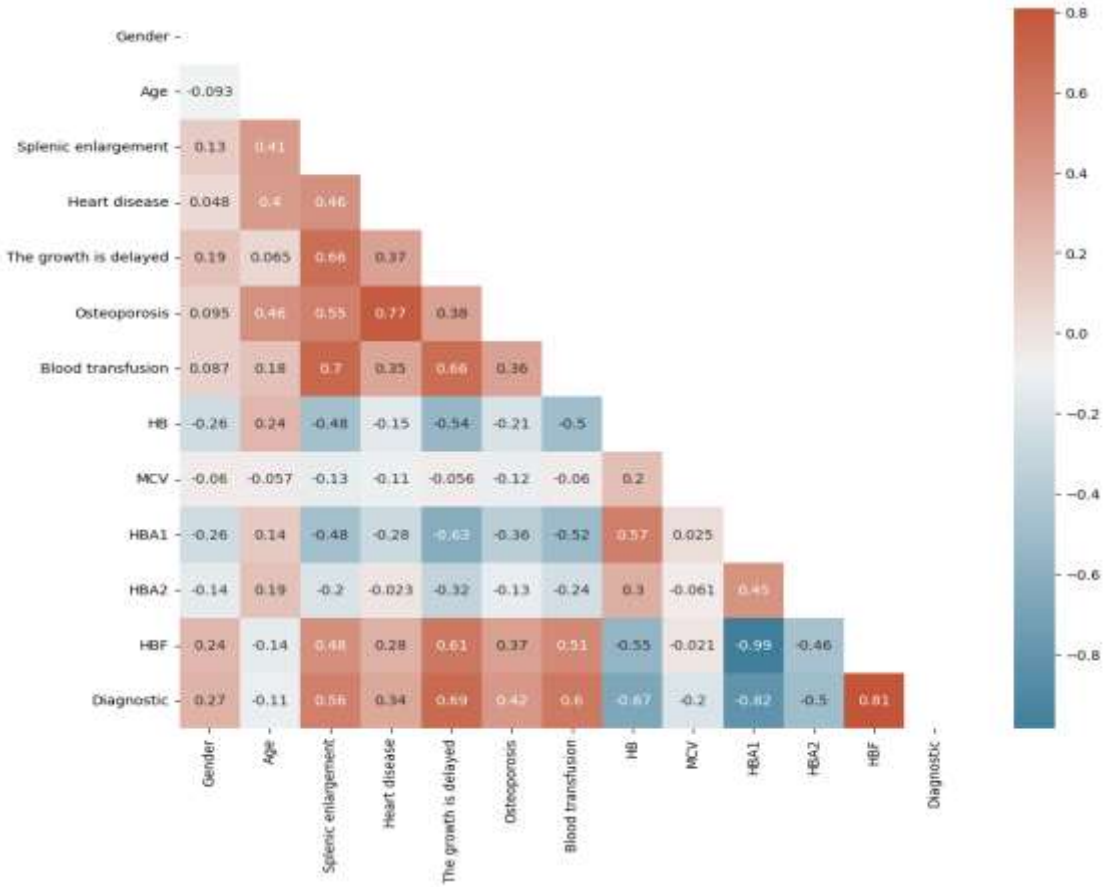


Figure 2 shows the Pearson correlation matrix for strongly and weakly correlated variables.

For example, the association between blood transfusion and splenomegaly is 0.7, whereas the correlation between blood transfusion and bone abnormalities is 0.38. Fetal hemoglobin (HBF) shows a strong positive association with prognosis (0.81), while hemoglobin A1 (HBA1) and mean corpuscular volume (MCV) show a strong and moderate significant association with prognosis, respectively (-0.82 and -0.67).

3. Methodology

3.1 Logistic Regression Model

Logistic regression (LR) is one of the most prominent statistical and data mining techniques used by statisticians and researchers to analyze and classify data sets with binary and multiple responses **(Maalouf, 2011)**. Logistic regression, often known as a logit model, assesses the connection between a dependent variable Y, which has only two potential values based on the occurrence or non-occurrence of an event, and explanatory factors that influence that phenomenon. Its popularity stems from the fact that it does not have to meet many of the requirements of linear regression and general linear models, such as a linear relationship between the dependent and independent variables, a normal or homogeneous distribution of the independent variables, or the use of variables measured using the metric system **(Ghosh et al., 2018)**.

The logistic model comes from the logistic function by defining z as a linear sum of the independent variables. This sum is substituted into the logistic function to obtain the probability result. **(M Gail, K. Krickeberg, 2010)** Where the logistic function represents the following formula:

$$f(z) = \frac{1}{1 + e^{-z}} \qquad (1)$$

Where Z= β0+ β1x1+ β2x2+… +βnxn
Where β0: the constant limit parameter represents or intercept. x1,x2,…,xn the Independent variables
β1, β2,…,βn: Regression coefficients represent. i=1,2,..n

$$P(Y = yi|x1, x2, ..xn) = \frac{1}{1 + e^{-(\beta 0 + \beta 1 x1 + \beta 2 x2 + \cdots \beta nxn)}} \quad (2)$$

Yi= 0,1
$P(Y = yi|x1, x2, ..xn)$ It represents the conditional probability of a particular event occurring based on a set of independent variables.

Odds Ratio: It is the ratio of the conditional probability of a specific event occurring to the conditional probability of the event not happening. As in the following formula:

$$\textbf{Odds} = \frac{P(Y=yi|x1,x2,..xn)}{1-P(Y=yi|x1,x2,..xn)} = \frac{\frac{1}{1+e^{-(\beta 0+\sum_{i=1}^{n} x_i\beta_i)}}}{1-\frac{1}{1+e^{-(\beta 0+\sum_{i=1}^{n} x_i\beta_i)}}} = e^{\beta 0+\sum_{i=1}^{n} x_i\beta_i} \quad (3)$$

The logit is calculated by taking the natural logarithm of the odds ratio, which represents the logistic regression model

$$\text{Logit}(P(Y=yi|x1,x2,..xn)) = \text{Ln}\left[\frac{\frac{1}{1+e^{-(\beta 0+\sum_{i=1}^{n} x_i\beta_i)}}}{1-\frac{1}{1+e^{-(\beta 0+\sum_{i=1}^{n} x_i\beta_i)}}}\right] \quad (4)$$

$$\text{Logit}(P(Y=yi|x1,x2,..xn)) = \beta 0 + \sum_{i=1}^{n} x_i\beta_i \quad (5)$$

3.2 K-Nearest Neighbors (KNN) Classifier

It is a supervised learning method that uses non-parametric estimation techniques for binary classification. The idea behind KNN is that if there are multiple observations from two separate data sets, each observation should be classified based on its class. This is achieved by calculating the distance between the new data point and all the data points in the training set using a suitable distance measure, which is known in the training set as K-nearest neighbor; hence, we need to determine the value of K before applying k-nearest neighbor, and then the distances are sorted in ascending order, with the nearest neighbors representing the distances with the lowest value. The classification procedure is then completed by determining the class to which the majority of points near the observation to be classified belong and then assigning the new data point to that class. The image below shows training samples for two different cycles (A and B). The KNN approach calculates the distance between the new observation (the test sample) and each observation from the training set **(Steinbach & Tan, 2009).** The graph 3 demonstrates the classification of the new observation as class B when K = 3 and class A when K = 6.

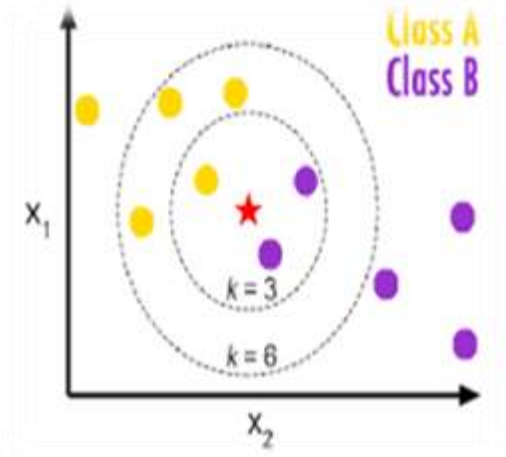


Figure 3 illustrates the KNN method based on (Rithesh, 2017).

3.2.1 Distance Measures Used in the KNN Algorithm are:

1. Makowski distance measure: This measure depends on the absolute difference between the two vectors (Xi, Yi) and its equation is as follows:

$$d(X,Y) = \sqrt[p]{\sum_{i=1}^{n}|xi - y_i|^p} \quad (6)$$

p denotes a positive value, where xi represents the value of i in vector X and yi represents the value of i in vector Y. When p=1, the Minkowski distance measure becomes the Manhattan distance measure; when p=2, it becomes the Euclidean distance measure; and when p=∞, it becomes the Chebyshev distance, as shown in the distances below:

First, Euclidean distance: Machine learning algorithms like KNN commonly use Euclidean distance as a distance measure, gauging the similarity or difference between data points based on their feature values. Euclidean distance represents the shortest distance between two vectors and is the square root of the sum of the squared differences between two vectors. Its equation is as follows:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (7)$$

where xi represents the observations of the independent variable to be classified, yi represents the observations of the independent variable's nearest neighbors, and n is the number of independent variables. **(Prakisya et al., 2021)**

Second, the city block distance: also known as the Manhattan distance, represents the sum of the absolute differences between two vectors. The Manhattan distance works very well with high-dimensional datasets. Since it is non-square, it does not inflate the differences between any of the variables. **(Gao & Li, 2020)**

$$\boldsymbol{d(x,y) = \sum_{i=1}^{n}|x_i - y_i|} \quad (\mathbf{8})$$

Third, Chebyshev distance: It is also referred to as the maximum distance, Lagrange distance, or chessboard distance. It is the largest distance for the absolute difference between the two vectors (X, Y). **(Prasath et al., 2017)** and its equation is as follows:

$$d(X,Y) = \max_i\{|xi - yi|\} \quad (9)$$

2- Hamming distance: This distance calculates the amount of mismatch between two vectors. This distance is usually used with nominal data, but can also be applied to numerical data. As in the following formula: -

$$d(X,Y) = \sum_{i=1}^{n} 1\ xi \neq yi \quad (10)$$

3- The Jaccard distance: to measure the differences between a set of observations, the Jaccard distance is used, as shown in the formula below:

$$d(X,Y) = \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}x_i^2 + \sum_{i=1}^{n}y_i^2 - \sum_{i=1}^{n}x_i y_i} \quad (11)$$

4- cosine distance: Cosine distance, also known as angle distance, is derived from cosine similarity which measures the angle between two vectors, because the cosine distance is obtained by subtracting the cosine similarity from one. **(Prasath et al., 2017)** As in the following equation:

$$d(X,Y) = 1 - \frac{\sum_{i=1}^{n}x_i y_i}{\sqrt{\sum_{i=1}^{n}x_i^2}\sqrt{\sum_{i=1}^{n}y_i^2}} \quad (\mathbf{12})$$

5- Correlation distance: Correlation distance modifies the Pearson correlation coefficient to make the distance between [0,1]. Pearson correlation coefficient measures the linear association between two variables and extends from [-1,1]. To make Pearson correlation coefficient a distance between [0,1]. **(Arora et al., 2022)** use the formula below:

$$d(X,Y) = \frac{1}{2}\left[1 - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}\right] \quad (13)$$

where $\bar{x} = \frac{1}{n}\sum_{i=1} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1} y_i$

4 Model Evaluation Measures

Binary classification models predict both positive and negative classes (0,1) several classification metrics that are calculated to evaluate the performance of binary classification models. Among the most commonly used metrics are accuracy, precision, recall, and F1-score, and these metrics can be calculated using a confusion matrix.

4.1 Confusion Matrix:

We compare the predicted values with the actual values in a confusion matrix that contains the correct and incorrect predicted values. The rows in the confusion matrix represent the actual classes, while the columns in the confusion matrix represent the classes predicted by the model. (Bakumenko & Elragal, 2022) As in the following table:

Table 4 shows the confusion matrix

		Prediction	
		0	1
Actual	0	TN True Positive	FP False positive
	1	FN False Negative	TP True Positive

TN: The number of views properly identified as negative.

FP: It represents the number of observations that are incorrectly classified as positive but are negative, which is considered a type I error.

FN: It refers to the number of observations that were incorrectly classified as negative but were positive, and is described as a type II error.

TP: This indicates the number of views correctly classified as positive.

Accuracy: is determined by dividing the number of correct predictions by the total number of correct and incorrect predictions produced by the model, as shown in the formula below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100 \quad (14)$$

Inaccuracy, often known as error, can be measured by subtracting the equation () from one, as shown in the following formula:

$$Error = 1 - Accuracy = 1 - \frac{TP + TN}{TP + TN + FP + FN}\ x\ 100 \quad (15)$$

Precision: Precision is defined as the ratio of true positive predictions divided by the total number of positive predictions. The higher the precision, the more accurate the model is at distinguishing between true positives and false positives. Precision is recommended when the goal is to reduce the number of false positive predictions. For example, if a model has high recall but low precision, it indicates that it identifies a majority of true positives while producing a large number of false positives. **(Sergue, 2020)** As in the following formula:

$$Precision = \frac{TP}{TP + FP} * 100 \quad (16)$$

Recall: Recall describes the model's ability to correctly detect all positive classes. If recall is high, it means that the model recognized most of the true positive classes (i.e., few false negatives). Recall plays a crucial role in minimizing the number of misclassified negative instances**. (Sergue, 2020)** As in the following formula:

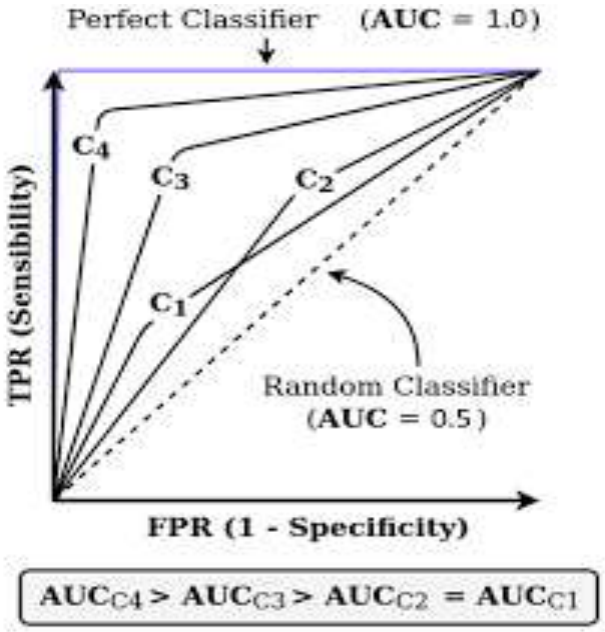
$$Recall = \frac{TP}{TP + FN} * 100 \quad (17)$$

F1-score: The F1 measure is the harmonic mean of accuracy and recall, aiming to get a weighted average of the two. **(Karlsson, 2017**) The F1 metric is defined below:

$$F1 - Score = 2 * \frac{Recall * Precision}{Recall + Precision} \quad (18)$$

4.2 The Area Under Curve (ROC curve)

The receiver operating characteristic (ROC) curve, as shown in the figure, calculates the area under the curve (AUC) by expressing the relationship between the true positive rate (TPR) and the false positive rate (FPR) at different thresholds. This curve helps us graphically analyze the performance of a classification model. Simply put, when AUC = 0.5, it indicates that the model is performing randomly, but if AUC < 0.5, it indicates that the model is performing worse than randomly, i.e. it is converting correct classifications to incorrect classifications and vice versa, where the closer the AUC value is to one, the better the model is performing, regardless of the threshold used to estimate the probability of a sample belonging to each class.(de Oliveira et al., 2021)**.**



5 Results and Discussion

After splitting the data into training and test sets, we started building the model using the training dataset. The confusion matrix (Figure) shows the actual classes and the predicted classes. The confusion matrix showed 194 patients with thalassemia (0 = 105 and 1 = 91). The confusion matrix represented thalassemia intermedia (0) and thalassemia major (1). Figure 5 below shows the results of applying the confusion matrix to the training data.

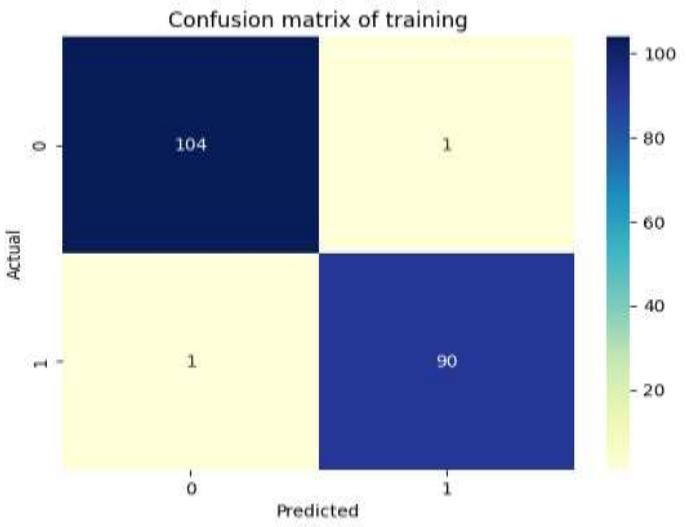


Figure 5 shows the confusion matrix for the logistic regression model on the training data

The model correctly classified 104 observations as having thalassemia intermedia and predicted that they would also have the condition. The model incorrectly classified one observation as having thalassemia major when it actually had thalassemia intermedia and incorrectly identified another observation as having thalassemia intermedia when it actually had thalassemia major. The value 90 represents the number of positive observations correctly classified as having thalassemia major, and the model also predicted that they would have thalassemia major. Figure 6 below shows the results of applying the confusion matrix to the testing data.

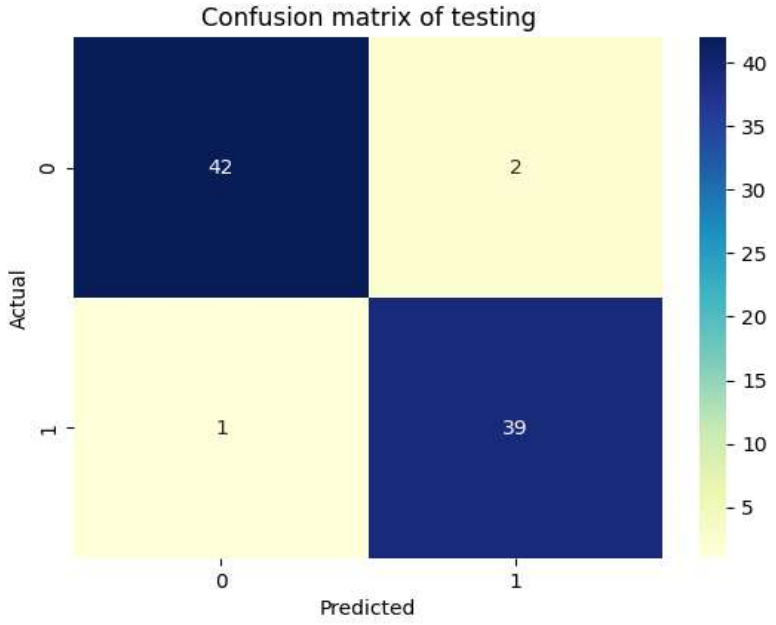


Figure 6 shows the confusion matrix for the logistic regression model on the training data

The number of observations correctly classified as having thalassemia intermedia is 42. The model incorrectly classified two values as having thalassemia major, but it correctly predicted them to have thalassemia intermedia. One instance was mistakenly identified as having thalassemia intermedia, but in reality, it was thalassemia major. The number of observations correctly classified as thalassemia major stands at 39. This indicates that the logistic regression model has perfect prediction ability. People often use precision and error metrics to evaluate the performance of a model. However, this metric does not provide a comprehensive understanding of the performance of the model. Our primary goal is to determine the efficiency of the model,

specifically its flexibility. We then use other, more effective metrics to evaluate the performance of the model and determine the degree to which our predictions match the actual values.

Table 5 shows the results of evaluating the performance of the logistic regression model distance

Models	**raining Accura**	**esting Accurac**	**Classes**	**Precision**	**Recall**	**F1-score**	**AUC**
Logistic Regression	0.98	0.96	0	0.96	0.98	0.97	
			1	0.97	0.95	0.96	

According to the findings in the table above, the logistic regression model's accuracy for training data is 98%, its accuracy for testing is 96%, the precision is 96%, the recall is 98%, and the F1 score is 97%. The second class has an F1 score of 0.96, recall of 0.95, and accuracy of 0.97. These findings show that the model performs well in discriminating between the two groups. Another way to measure the performance of a model is the receiver operator characteristic (ROC). The vertical axis shows the true positive rate, while the horizontal axis shows the false positive rate. We determine the validity of a model by calculating the area under the ROC curve (AUC). The LR model for the test set is represented by the AUC value (0.98). As shown in the figure 7

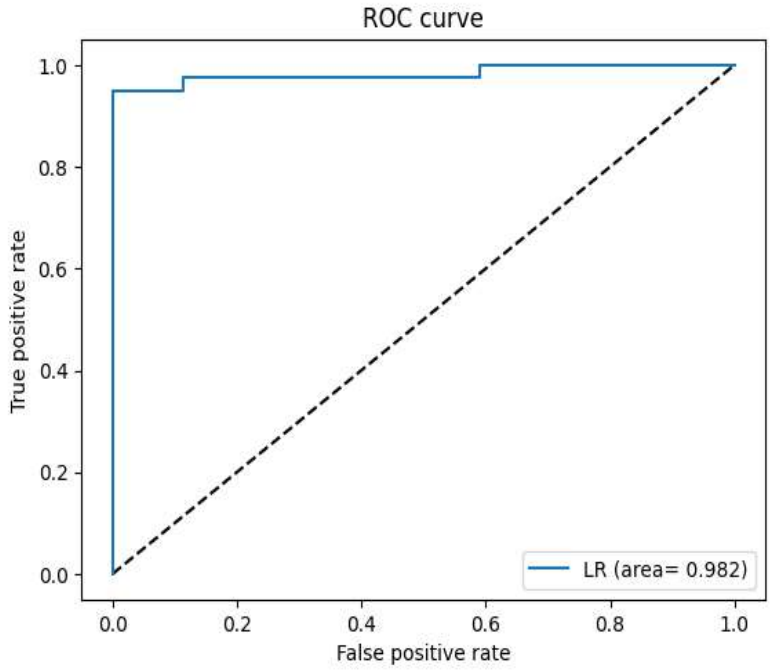


Figure 7 Receiver operating characteristic curve for logistic regression model

KNN is a basic the supervised learning algorithm that predicts the appropriate class for new data based on the majority of its nearest neighbors. As a result, the parameter K is important in evaluating this classifier. To determine the optimal value of k, we used 5-fold cross-validation. The graph 8 shows the accuracy of a k-NN classifier using cross-validation. The vertical axis represents the model accuracy, while the horizontal axis represents the k values, which range from 1 to 15. At each value of k, the Chebyshev distance was used to determine performance.

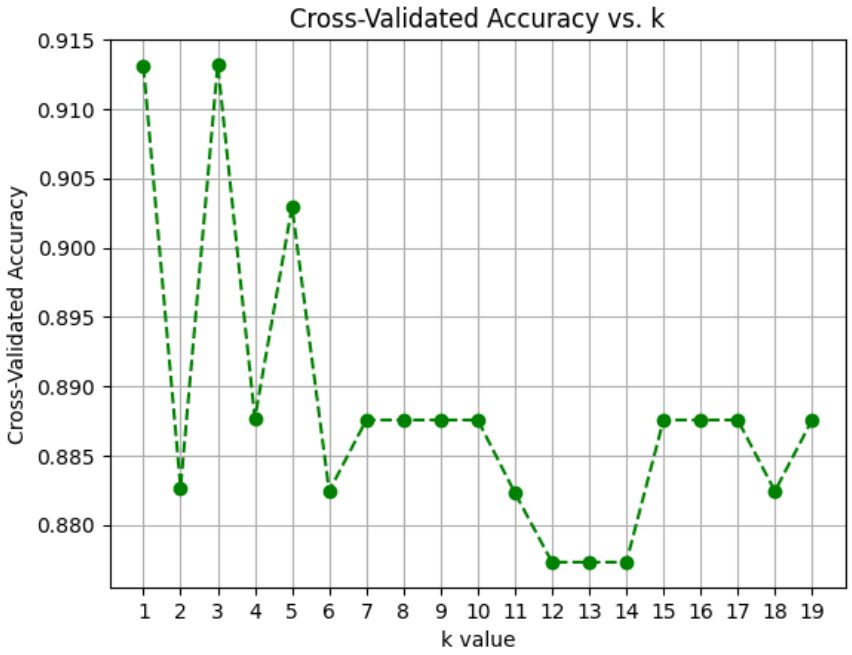


Figure 8 shows the cross-validation accuracy of the KNN classifier when using Chebyshev distance across different values of k

The results indicate that the best accuracy was obtained when the k values were 1, 3, which is about 91%. After that, we notice that the accuracy fluctuates with increasing k values. This indicates that the model does not improve after k values are 1, 3. Below is the training and testing matrix for the nearest neighbor classifier using Chebyshev distance to see how many predictions are correct and incorrect.

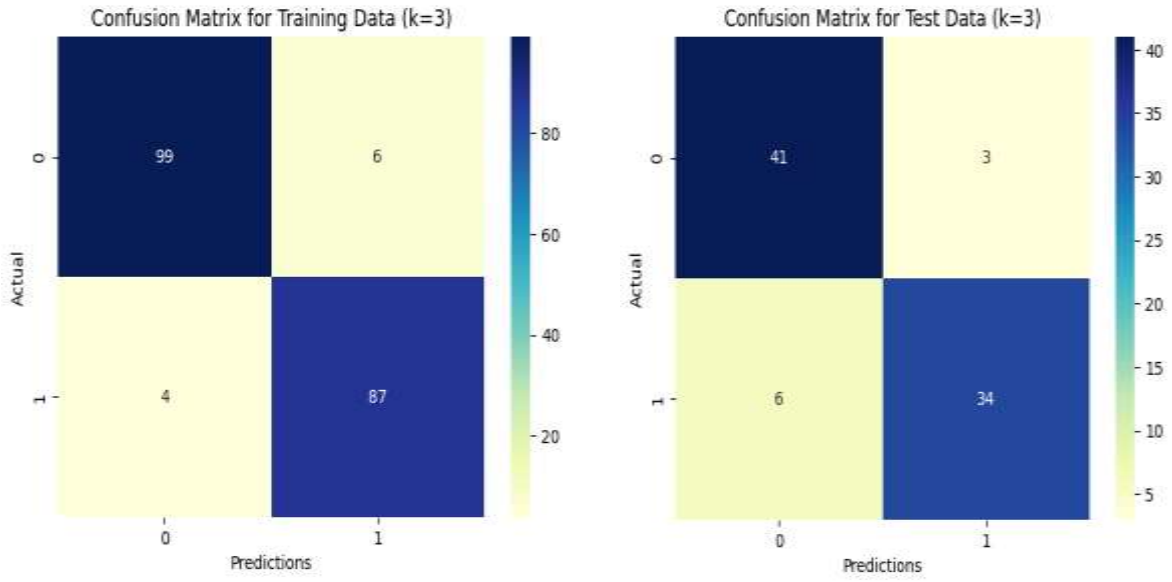


Figure 9 shows the accuracy of the KNN classifier when using Chebyshev distance across different values of k.

The figure shows the confusion matrix for the training data. The value 99 represents the number of observations that were correctly classified as having thalassemia intermedia, while the value 87 represents the number of observations that were correctly classified as having thalassemia major. The model made 10 errors. As for the test matrix, the model correctly predicted 75 values and made 9 errors. This indicates that the KNN classifier performs well when using Chebyshev distance.
The table below displays the classification report generated when the Chebyshev distance is employed with the KNN classifier. Table 6 shows training accuracy, test accuracy, and performance assessment metrics (precision, recall, and F1-score) for the two categories (0, 1).

Table 6 shows the performance evaluation results of the KNN model using Manhattan distance.

Models	**raining Accura**	**esting Accurac**	**Classes**	**Precision**	**Recall**	**F1-score**	**AUC**
NN with Chebysh K=3	0.95	0.89	0	0.87	0.93	0.90	
			1	0.91	0.85	0.88	

The results indicate that the model achieved 95% accuracy in training and 89% in testing. For thalassemia intermedia, the precision was 87%, recall was 93%, and F1 was 90%. For thalassemia major, the precision was 91%, recall was 85%, and F1 was 88%. The figure below shows the receiver operating characteristic curve to distinguish between the true positive rate and the false positive rate at different thresholds. The area under the curve (AUC) value is 95%, which indicates that the can distinguish between thalassemia intermedia and thalassemia major patients with 95% accuracy.

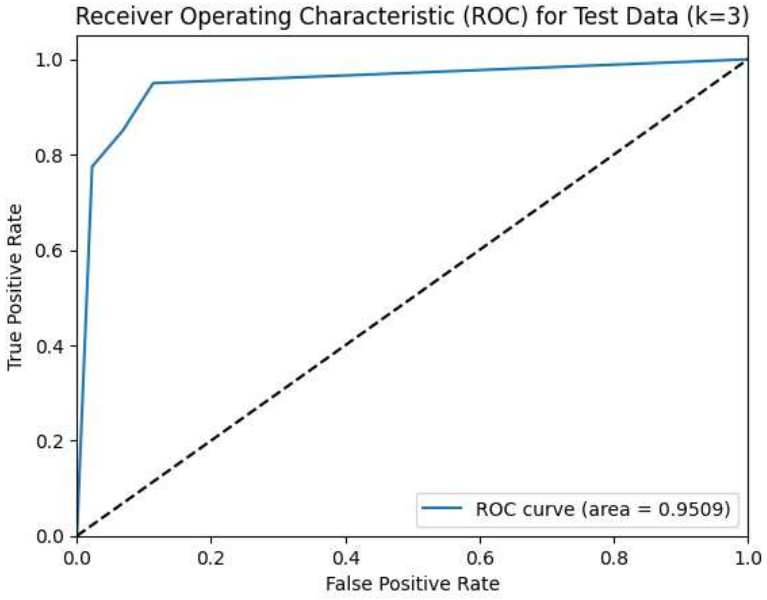


Figure 10 shows the receiver operating characteristic curve of a KNN class using Chebyshev distance

The graph 11 shows the accuracy of a k-NN classifier using K-fold cross-validation. The vertical axis represents the model accuracy, while the horizontal axis represents the k values, which range from 1 to 15. For each value of k, the Manhattan distance was used to determine the model performance.

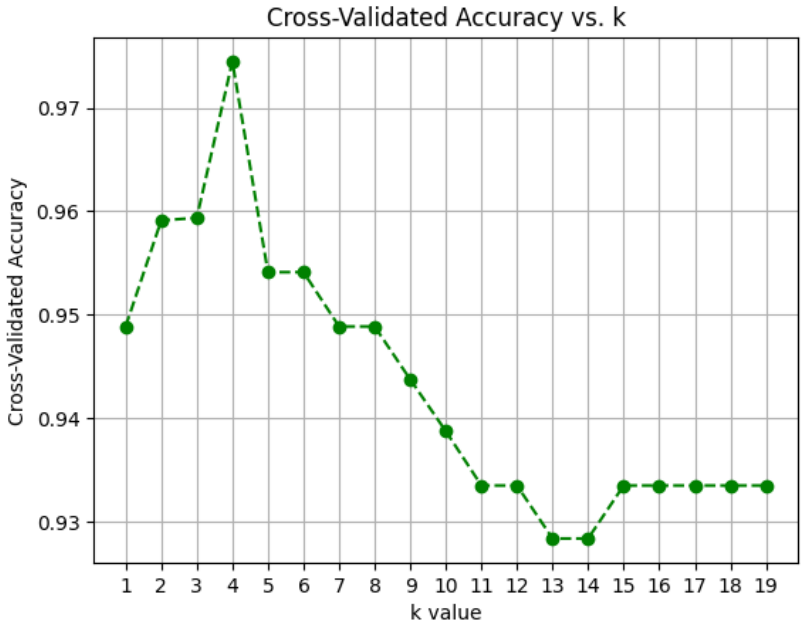


Figure 11 shows the accuracy of the KNN classifier when using Manhattan distance across different values of k

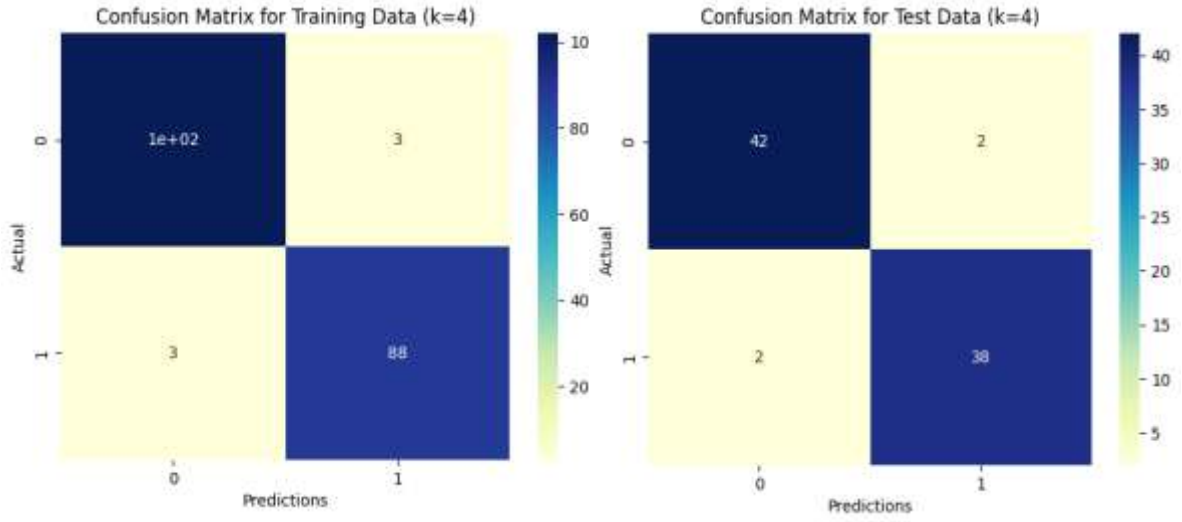


Figure 12 shows the accuracy of the KNN classifier when using Manhattan distance across different values of k

The confusion matrix displays predicted and corrected values for training and test samples. The training matrix shows that the model accurately predicted 188 out of 194 values while misprediction 6 values. The test matrix yielded 80 accurate predictions and 4 incorrect predictions.

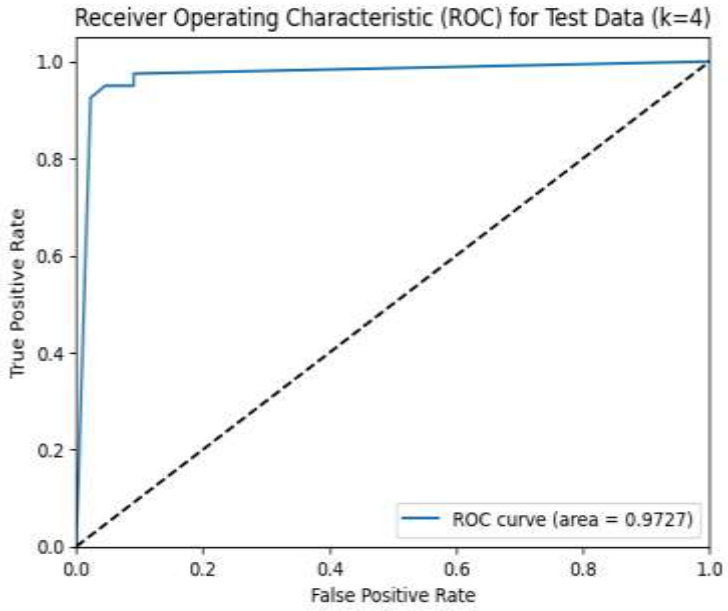


Figure 13 shows the receiver operating characteristic curve of a KNN class using Manhattan distance

The graph 13 above shows the ROC curve, which represents the relationship between the true positive rate and the false positive rate across different thresholds, where the main diagonal represents the performance of the random classifier. The closer the curve is to the upper left, the better the performance. The AUC value is 0.97, which indicates that the model has high reliability in binary classification.

The graph 13 above indicates that the accuracy value is initially approximately 95% when the value of k is one, then it starts to rise and the accuracy stabilizes at 96% in the case of k 2.3, while the best accuracy for precision is 97% in the case of k = 4, then after that the accuracy starts to gradually decrease as the value of k increases, and this indicates that the model does not improve after using the value of k = 4.

Table 7 below shows the resulting classification report when using Manhattan distance with the KNN classifier. The table shows training accuracy, test accuracy, and performance evaluation metrics precision, recall, and F1 score for the two classes (0, 1).

Table 7 shows the results of the performance evaluation criteria for KNN mode

Models	**raining Accurac**	**esting Accurac**	**Classes**	**Precision**	**Recall**	**F1-score**	**AUC**
N with Manhat K=4	0.97	0.95	0	0.96	0.96	0.96	
			1	0.95	0.95	0.95	

The results indicate that the model achieved 97% accuracy in training and 95% in testing. For the first class, the accuracy was 95%, the recall was 95%, and the F1 was 95%. For the second class, the accuracy was 95%, recall was 95%, and F1 was 95%. The figure below shows the receiver operating curve for distinguishing between the true positive rate and the false positive rate at different thresholds. The area under the curve (AUC) value is 97%, indicating that the model is able to distinguish between thalassemia intermedia and thalassemia major patients with 97% accuracy.

Above, we examined the logistic regression model for training and testing data, as well as the closest neighbor model when employing Chebyshev and Manhattan distances for training and testing, depending on the performance assessment criteria utilized. The table below compares the two models (logistic regression and the closest neighbor technique).

Depending on the performance assessment criteria utilized, we looked at the logistic regression model for training and test data, as well as the closest neighbor model for training and testing when Chebyshev and Manhattan distances were employed. Table 8 below compares the two models (logistic regression and the closest neighbor approach). The table includes test accuracy, precision, recall, F1 score, and AUC value. We rely only on the test set since determining the optimum model requires test data.

The Table 8 shows a comparison of the performance of the methods used

Models	**esting Accurac**	**Classes**	**Precision**	**Recall**	**F1-score**	**AUC**
Logistic Regression	0.96	0	0.96	0.98	0.97	0.982
		1	0.97	0.95	0.96	
NN with Chebyshe K=3	0.89	0	0.87	0.93	0.90	0.950
		1	0.91	0.85	0.88	
NN with Manhatta K=4	0.95	0	0.96	0.96	0.96	0.972
		1	0.95	0.95	0.95	

Since the logistic regression model has a logistic regression model accuracy of 96%, the KNN model accuracy in case of using Chebyshev distance is 89%, and the KNN accuracy in case of using Manhattan distance is 95%, the preference is given to the logistic regression model. The table just presented provides a summary of these results. In addition, when we consider other performance evaluation criteria, we find that the logistic regression model is better than other models in terms of accuracy (96%), recall (98%), and F1 score (97%) in the first category. As for the second category, the accuracy is 97%, recall is 95%, and F1 score is 96%. Since the area under the curve (AUC) value of the logistic regression model is 0.982, it is clear that this model is better than other models as well. In the case of comparing the nearest neighbor model using Manhattan distance with Chebyshev distance. The nearest neighbor model with Manhattan distance was better based on the performance evaluation results.

6 Conclusions

Given the importance of thalassemia and the number of deaths and health problems it causes, especially in thalassemia major, it has been shown that logistic regression is more effective than the K-nearest neighbor algorithm in classifying thalassemia patients, especially those with thalassemia major. The study showed that the type of distance used in the K-nearest neighbor algorithm, whether "Manhattan" or "Chebyshev", has a significant impact on the accuracy of predictions, with the highest accuracy reaching 95% when K = 4. It was also shown that the difference between distance calculation methods and the K value plays a major role in improving the classification results, as it was determined that the optimal value for K is 4, which led to improving the accuracy of predictions. The researcher suggests increasing the data size, as it is possible to improve the accuracy of models by increasing the data size. In addition, the researcher recommends using other artificial intelligence techniques, especially neural networks, to verify any additional improvements.

References

1. Arora, I., Khanduja, N., & Bansal, M. (2022). Effect of Distance Metric and Feature Scaling on KNN Algorithm while Classifying X-rays. CEUR Workshop Proceedings, 3176, 61–75.
2. Bakumenko, A., & Elragal, A. (2022). Detecting Anomalies in Financial Data Using Machine Learning Algorithms. Systems,

10(5). https://doi.org/10.3390/systems10050130
3. Borah, M. S., Bhuyan, B. P., Pathak, M. S., & Bhattacharya, P. K. (2018). Machine learning in predicting hemoglobin variants. International Journal of Machine Learning and Computing, 8(2), 140–143. https://doi.org/10.18178/ijmlc.2018.8.2.677
4. de Oliveira, N. R., Pisa, P. S., Lopez, M. A., de Medeiros, D. S. V., & Mattos, D. M. F. (2021). Identifying fake news on social networks based on natural language processing: Trends and challenges. Information (Switzerland), 12(1), 1–32. https://doi.org/10.3390/info12010038
5. Gao, X., & Li, G. (2020). A KNN Model Based on Manhattan Distance to Identify the SNARE Proteins. IEEE Access, 8, 112922–112931. https://doi.org/10.1109/ACCESS.2020.3003086
6. Ghosh, J., Li, Y., & Mitra, R. (2018). On the use of cauchy prior distributions. Bayesian Analysis, 13(2), 359–383.
7. Hartini, S., & Rustam, Z. (2019). Hierarchical clustering algorithm based on density peaks using kernel function for thalassemia classification. Journal of Physics: Conference Series, 1417(1), 12016.
8. Karlsson, S. (2017). Using semantic folding with TextRank for automatic summarization. 58.
9. M Gail, K. Krickeberg, J. M. S. (2010). Statistics for Biology and Health. In Media.
10. Maalouf, M. (2011). Logistic regression in data analysis: An overview. International Journal of Data Analysis Techniques and Strategies, 3(3), 281–299. https://doi.org/10.1504/IJDATS.2011.041335
11. Paokanta, P., Ceccarelli, M., & Srichairatanakool, S. (2010). The effeciency of data types for classification performance of machine learning techniques for screening β-Thalassemia. 2010 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies, ISABEL 2010, 1–4. https://doi.org/10.1109/ISABEL.2010.5702769
12. Prakisya, N. P. T., Liantoni, F., Hatta, P., Aristyagama, Y. H., & Setiawan, A. (2021). Utilization of K-nearest neighbor algorithm for classification of white blood cells in AML M4, M5, and M7. Open Engineering, 11(1), 662–668. https://doi.org/10.1515/eng-2021-0065
13. Prasath, V. B. S., Alfeilat, H. A. A., Hassanat, A. B. A., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., & Salman, H. S. E. (2017). Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier -- A Review. 1–39. https://doi.org/10.1089/big.2018.0175
14. Rithesh, R. N. (2017). SVM-KNN: A Novel Approach to Classification Based on SVM and KNN. International Research Journal of Computer Science, 4(8), 43–49. https://doi.org/10.26562/irjcs.2017.aucs10088
15. Sergue, M. (2020). Customer Churn Analysis and Prediction using Machine Learning for a B2B SaaS company. www.kth.se/sci
16. Steinbach, M., & Tan, P.-N. (2009). kNN: k-nearest neighbors. *The top ten algorithms in data mining*, 151-162 .
17. Yousefian, F., Banirostam, T., & AzarKeivan, A. (2017). Prediction of Mellitus Diabetes in Patients with Beta-thalassemia using Radial Basis Network, and k-Nearest Neighbor based on Zafar Thalassemia Datasets. Diabetes, 19, 20.

تقييم نماذج تعلم الالة لتشخيص مرض الثلاسيميا باستخدام الانحدار اللوجستي وخوارزمية الجار الأقرب

محمد فارس علي[1] وحذيفة حازم طه[2]

[2,1]دراسات عليا ، قسم الإحصاء والمعلوماتية، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق.

قسم بحوث العمليات والتقنيات الذكائية ، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق.

الخلاصة : الثلاسيميا مرض وراثي ينتقل من الوالدين إلى الأبناء عندما يكون الوالدان ناقلين للطفرة الوراثية. ويؤدي هذا التغير إلى انخفاض في عدد ونوعية وحالة الصفائح الدموية الحمراء وزيادة في معدل تلف الصفائح الدموية الحمراء، مما يؤدي إلى تراكم الحديد في الجسم ونقص الهيموجلوبين في الدم. يهدف هذا البحث إلى تطوير نموذج للتنبؤ بالثلاسيميا باستخدام اسلوبين من اساليب الذكاء الاصطناعي وهما الانحدار اللوجستي وخوارزمية الجار الاقرب بناءً على معايير تقييم اداء النماذج: Accuracy, Precision, Recall, F1−score, AUC. تم الحصول على البيانات من مستشفى الحدباء التخصصي في الموصل. تضمنت مجموعة البيانات 280 مشاهدة، منها 149 (53.21%) كانت ثلاسيميا متوسطة 131 (46.78%) كانت ثلاسيميا كبرى، تم تقسيم البيانات إلى 70% للتدريب و 30% للاختبار. وأظهرت النتائج التجريبية أن نموذج الانحدار اللوجستي كان أفضل أداءً من خوارزمية أقرب جار بدقة 96% وتذكر 98% ودرجة F1− 97% في فئة الثلاسيميا المتوسطة، بينما كان له دقة 97% وتذكر 95% ودرجة 96% في فئة الثلاسيميا الكبرى، مما يدل على أن الانحدار اللوجستي كان جيدًا في التمييز بين هاتين الفئتين. وقد ثبت أن الانحدار اللوجستي أكثر فعالية من خوارزمية أقرب جار K في تصنيف مرضى الثلاسيميا، وخاصة المصابين بالثلاسيميا الكبرى. وأظهرت الدراسة أن نوع المسافة المستخدمة في خوارزمية الجار الاقرب سواء "مانهاتن" أو "تشيبيشيف" لها تأثير كبير على دقة التنبؤات حيث تصل أعلى دقة إلى 95% عندما تكون K=4. كما تبين أن الفرق بين طرق حساب المسافة وقيمة K يلعب دوراً كبيراً في تحسين نتائج التصنيف حيث تم تحديد أن القيمة المثلى لـ K هي 4 مما أدى إلى تحسين دقة التنبؤات. ويقترح الباحث زيادة حجم البيانات حيث من الممكن تحسين دقة النماذج بزيادة حجم البيانات. بالإضافة إلى ذلك يوصي الباحث باستخدام تقنيات الذكاء الاصطناعي الأخرى وخاصة الشبكات العصبية للتحقق من أي تحسينات إضافية.

الكلمات المفتاحية: التعلم الآلي, التنبؤ بمرض الثلاسيميا ,نموذج KNN