1st International Conference on Artificial Intelligence in the Era of Digital Transformation
Event Place: Tbilisi,Georgia
www.Aicnf.ir
اولین کنفرانس بین المللی هوش مصنوعی در عصر تحول دیجیتال | گرجستان
1st International Conference on Artificial Intelligence in the Era of Digital Transformation
PUBLISH IN JOURNALS — INTERNATIONAL CERTIFICATION

Diagnosis of Infertility in Women with Thalassemia Using a Hybrid Particle Swarm Optimization and Multilayer Perceptron Approach

Fatemeh Hoseinkhani[1], Atefeh Abedi[2]
[1]Ph.D Assiatant in Artificial Intelligence, Qazvin University of Medical Sciences, Qazvin, Iran.
[2]M.Sc in Medical Engineering, Qazvin Islamic Azad University, Qazvin, Iran.

Abstract

One of the most common diseases in today's world is thalassemia, and its prevalence is increasing globally every year. The use data mining techniques to develop predictive models for identifying individuals at risk of thalassemia is highly beneficial in reducing complications associated with the disease. These techniques utilize statistical methods and artificial intelligence to identify patterns and relationships among variables.In this study, a data mining approach based on the C4.5 decision tree, multilayer perceptron (MLP) neural network, and a hybrid model combining particle swarm optimization (PSO) with MLP was used to analyze criteria and predict infertility in thalassemia patients. A total of 55 criteria were considered for thalassemia evaluation, which were used as input neurons in the neural network. The results indicate that the root mean square error (RMSE) for the C4.5 decision tree is 0.0216, for the MLP neural network is 0.0248, and for the hybrid PSO-MLP model is 0.0276. The classification accuracy for the C4.5 decision tree, MLP neural network, and hybrid PSO-MLP model on test data is 100%, 96.8254%, and 100%, respectively. It can be concluded that, for the given dataset and based on the clustering performed on the data, the C4.5 decision tree and the hybrid PSO-MLP model outperform the standalone MLP neural network in terms of predictive performance.

Keywords: Thalassemia, Infertility, Decision Tree, Data Mining, Particle Swarm Optimization.

1- Introduction

Thalassemia is a genetic disorder in which hemoglobin loses its normal structure, leading to the production of non-functional hemoglobin. Thalassemia is classified into two types: alpha and beta. Alpha thalassemia is usually silent and is sometimes referred to as beta thalassemia minor. Beta thalassemia is one of the most common quantitative hemoglobin disorders worldwide, particularly in Iran. Based on gene inheritance patterns, clinical symptoms, and the need for blood transfusions, thalassemia is categorized into four groups: Minima, Minor, Intermedia, and Major. Thalassemia intermedia refers to patients with milder anemia, who do not require regular blood transfusions, as their hemoglobin levels can be maintained at an optimal level without transfusions. However, in some cases, blood transfusions may be needed to prevent severe bone deformities. In contrast, thalassemia major patients suffer from severe anemia, requiring regular blood transfusions every 4–5 weeks before the age of one to maintain hemoglobin levels above 10 g/dL. This treatment regimen allows patients to lead a normal life, prevents progressive bone marrow expansion, and reduces the risk of aesthetic deformities and osteoporosis. However, hemosiderosis is an inevitable consequence of prolonged blood transfusion (Behram et al., 2008). Patients with beta-thalassemia major suffer from severe transfusion-dependent anemia, leading to iron deposition in endocrine glands, which results in infertility. These patients also experience delayed sexual development, and only a few cases of pregnancy have been reported (Tampakoudis et al., 1997 and Surbek et al. 1996). These patients experience poor sexual development, and only a few cases of pregnancy have been reported among them. Limited information is available regarding pregnancy and fertility in Mediterranean patients with beta-thalassemia major (Gibbons et al. 2001). Before the introduction of new treatments in the 1960s, pregnancy was only observed in cases of thalassemia intermedia. Due to hypogonadism resulting from iron deposition, ovulation disorders, and impaired fetal growth within the uterus, complications arise. Consequently, reduced hemoglobin levels and fetal hypoxia have been associated with preterm and premature births (Aessopos et al.1999). The goal of this research is to develop a model that integrates neural networks with particle swarm optimization (PSO) to identify and predict infertility risk factors in thalassemia patients before the onset of the condition. This early prediction aims to help prevent infertility in affected individuals. Additionally, another key objective of this study is to detect infertility in thalassemia patients with high accuracy and speed by analyzing specific parameters fed into the neural network.

2- Related Work

have relied on citation analysis methods with a statistical approach. Given the large volume of medical data, the use of modern data mining techniques has gained significant attention in recent research. Studies based on multilayer perceptron (MLP) neural networks and decision trees have demonstrated the superior performance of the C4.5 decision tree in A review of the research literature indicates that studies on predicting thalassemia at the global level are limited, and most predicting disease onset, identifying influential factors, and generating predictive and diagnostic rules. This advantage makes C4.5 a preferred method compared to algorithms such as logistic regression, classification trees, and other decision tree models. Comparative analyses between MLP neural networks and the C4.5 decision tree suggest that both methods produce acceptable and closely comparable results, indicating their efficiency and reliability in disease diagnosis. Among data mining techniques, decision trees and three-layer perceptron neural networks have proven useful in medical predictions and disease diagnosis. Furthermore, studies based on MLP neural networks have shown that this algorithm performs better than other algorithms examined in this field.

The present study is applied in terms of its objective. Regarding research methodology, it is a survey-based, descriptive-analytical study, and in terms of data collection, it follows a survey approach. Additionally, based on its approach and nature, it is a qualitative-quantitative study. At the initial stage of the research, criteria identification will be conducted through interviews with experts and specialists at Adult Thalassemia Clinic, as well as by reviewing academic theses and research articles. This process aims to gather relevant indicators for analyzing the relationship between female infertility and thalassemia, which will be used to design a dataset. Once the dataset structure is defined, individual files for each patient will be created based on the specified dataset fields, and the dataset will be completed accordingly.

Thalassemia is a hereditary blood disorder characterized by the body’s inability to produce sufficient hemoglobin, leading to anemia and other health complications. Early and accurate diagnosis is crucial for effective management and prevention of severe outcomes. In recent years, data mining and machine learning techniques have been increasingly applied to enhance the diagnostic processes for thalassemia.

1. Detection of β-Thalassemia Carriers Using Data Mining Techniques

A study published in the Sri Lanka Journal of Applied Statistics (2018) aimed to develop a time-efficient model to detect β-thalassemia carriers. The researchers employed data mining techniques to analyze measurable blood features, reducing the time required for decision-making and potentially decreasing the need for expensive and time-consuming tests (AlAgha et al.2018).

2. Applications of Artificial Intelligence in Thalassemia

A comprehensive review by Alzubaidi et al. (2023) examined the effectiveness of artificial intelligence (AI) in the diagnosis and classification of thalassemia. The study highlighted the use of various AI techniques to distinguish thalassemia from other causes of microcytic hypochromic anemia, particularly iron deficiency anemia. The authors

emphasized the importance of AI in aiding the diagnostic process, potentially reducing the need for more invasive and costly tests (Ferih et al. 2023).

3. Predicting Thalassemia Using Feature Selection Techniques

A study published in 2022 focused on predicting thalassemia using feature selection techniques. The researchers conducted a systematic review of AI-based and machine learning-based thalassemia diagnostic methods, analyzing various datasets, preprocessing methods, and classification techniques. The study provided insights into the effectiveness of different feature selection methods in improving diagnostic accuracy (Saleem et al. 2023).

4. Automated Diagnosis of Thalassemia Based on Data Mining Classifiers

An earlier study explored the use of data mining classifiers for the automated diagnosis of thalassemia. The researchers applied various classification algorithms to predict the risk of thalassemia, aiming to improve the decision-making process within clinical settings. The study concluded that using classifiers such as the multi-layer perceptron could enhance the accuracy of thalassemia predictions (El-Halees and Alshami. 2012). In another study, researchers developed a machine learning-based framework using red blood cell parameters to predict the α+-thalassemia trait. The study demonstrated that the model achieved an accuracy of 80.77% and a sensitivity of 70.59% in predicting the trait, indicating the potential of neural network-based approaches in large-scale thalassemia screening (Phirom et al. 2022).

5. A Comparative Review on Machine and Deep Learning Techniques

A recent review compared various machine and deep learning techniques applied to thalassemia diagnosis. The authors discussed different data mining methods used to identify beta-thalassemia carriers among asymptomatic individuals. The study highlighted the potential of combining data balancing methodologies, such as SMOTE and ADA, with machine learning algorithms to achieve high diagnostic accuracy (Kaur and Garg.2025).

6. An Application of Machine Learning to Thalassemia Diagnosis

A study proposed two modeling methods to predict whether patients have Mediterranean anemia (a form of thalassemia). The first method involved using Principal Component Analysis (PCA) followed by logistic regression modeling, while the second method utilized Partial Least Squares Regression (PLS). The experimental results demonstrated good predictive performance for both models, indicating their potential utility in clinical settings.

Comparison of Studies on Thalassemia Diagnosis Using Data Mining Techniques(Liu.2024).

The following table summarizes the key aspects of the discussed studies:

Table 1- An overview of existing methods of data minig for diagnosis of thalassemia

Study Title	Year	Objective	Methodology	Key Findings
Detection of β-Thalassemia Carriers Using Data Mining Techniques	2018	Develop a time-efficient model for detection	Data mining techniques analyzing measurable blood features	Reduced decision-making time, potential decrease in need for expensive tests
Applications of Artificial Intelligence in Thalassemia	2023	Review AI effectiveness in diagnosis and classification	Systematic review of AI techniques distinguishing thalassemia from other anemias	AI aids diagnostic process, potentially reducing need for invasive and costly tests
Predicting Thalassemia Using Feature Selection Techniques	2023	Predict thalassemia using feature selection	Systematic review of AI-based and ML-based diagnostic methods	Insights into effectiveness of different feature selection methods in improving accuracy
Automated Diagnosis of Thalassemia Based on Data Mining Classifiers	2012	Automated diagnosis using data mining classifiers	Application of various classification algorithms	Multi-layer perceptron enhances accuracy of thalassemia predictions
A Comparative Review on Machine and Deep Learning Techniques	2025	Compare machine and deep learning techniques	Review of data mining methods identifying beta-thalassemia carriers	Combining data balancing methodologies with ML algorithms achieves high diagnostic accuracy
An Application of Machine Learning to Thalassemia Diagnosis	2024	Predict Mediterranean anemia using ML models	PCA followed by logistic regression; Partial Least Squares Regression	Both models demonstrate good predictive performance, indicating potential utility in clinical settings

These studies collectively underscore the significant advancements in applying data mining and machine learning techniques to the diagnosis of thalassemia. The integration of these technologies holds promise for more accurate, efficient, and early detection, ultimately contributing to better patient outcomes and management strategies.

3- Proposed Method

Research methodology is a systematic process that always begins with a research question or problem, aiming to provide a scientific response to the stated issue. Achieving research objectives is only possible when the search for knowledge and research methodology is conducted correctly. Therefore, adopting a scientific approach is the only way to attain valid and reliable research findings. In other words, research methodology is a set of rules, tools, and structured methods used to examine facts, uncover unknowns, and find solutions to problems. The selection of a research method depends on the objectives, nature of the research topic, and available resources. Thus, a research method can only be determined once the nature, scope, and objectives of the study are clearly defined.

The present study is applied in terms of its objective. Regarding research methodology, it is a survey-based, descriptive-analytical study, and in terms of data collection, it follows a survey approach. Additionally, based on its approach and nature, it is a qualitative-quantitative study. At the initial stage of the research, criteria identification will be conducted through interviews with experts and specialists at Adult Thalassemia Clinic, as well as by reviewing academic theses and research articles. This process aims to gather relevant indicators for analyzing the relationship between female infertility and thalassemia, which will be used to design a dataset. Once the dataset structure is defined, individual files for each patient will be created based on the specified dataset fields, and the dataset will be completed accordingly.

The identified 65 fields in your dataset encompass a wide range of medical, genetic, and biochemical factors relevant to infertility assessment in women with thalassemia. Following is an organized and translated version of the fields: Name, Family, Sex, Type of thalassemia, Age at marriage, Spouse's health status (appears twice, possibly an error), Spontaneous pregnancy, Pregnancy with medication, Miscarriage, Number of IVF attempts, Beta thalassemia genetics, Alpha thalassemia genetics, XMN1 (genetic marker), Blood group, Rh factor (blood RH), Height, Weight, Age of first blood transfusion, Blood transfusion frequency, Spleen removal (splenectomy), Age at splenectomy, Puberty status, Treatment History, Type of iron chelation therapy, Hydroxyurea consumption, Liver iron concentration (LIC), Cardiac MRI (cardMRI), CT2MS (cardiac T2 MRI scan), Echocardiography EF (EchoEF), Echocardiography PAP (EchoPAP), Liver MRI (liverMRI), LT2MS (liver T2 MRI scan), Bone density – femur, Bone density – spine, White Blood Cell count (WBC), Hemoglobin (Hb), Platelet count, Fasting blood sugar (FBS), Uric acid, Blood Urea Nitrogen (BUN), Creatinine (Cr), Liver enzymes (SGOT, SGPT), Ferritin (iron storage marker), Calcium (Ca), Luteinizing hormone (LH), Follicle-stimulating hormone (FSH), Thyroid-stimulating hormone (TSH), Parathyroid hormone (PTH), Vitamin D (VitD), Estradiol, Progesterone (Porogestron), Total testosterone (Totestron), HBA (Hemoglobin A), HBF (Fetal hemoglobin), HBS (Sickle cell hemoglobin, if applicable), HBH (Hemoglobin H), HBA2 (Hemoglobin A2), HCV antibodies (HCVab, Hepatitis C test), Hepatitis (general status), Alloantibodies (AlloAb), Autoantibodies (AutoAb), Ovarian ultrasound, Uterine ultrasound, Date of birth, Description of medical procedures and interventions. These fields provide a comprehensive dataset for analyzing infertility in women with thalassemia, considering genetic, biochemical, hormonal, and clinical factors. If you need further categorization or refinement, let me know!

After data collection, data mining techniques will be used to measure, evaluate, and identify infertility in women with thalassemia. First, the K-means clustering algorithm will be applied to classify the impact of criteria into two clusters: infertile and non-infertile. Next, the criteria will be analyzed using the C4.5 decision tree algorithm, the multilayer perceptron (MLP) neural network, and a hybrid model combining MLP with the particle swarm optimization (PSO) algorithm. The C4.5 decision tree algorithm will be used to extract decision rules and determine the relationship between female infertility and thalassemia. Hybrid models of MLP and PSO algorithm are, as their name suggests, a combination of computational intelligence approaches that aim to leverage each other's strengths for problem-solving and solution optimization. In such hybrid systems, different techniques are either integrated or used alongside each other. Typically, embedding one system within another helps compensate for their weaknesses, enhance their strengths, or achieve both benefits simultaneously. By combining these methods, hybrid systems gain significant advantages. Specifically, the high-speed parameter estimation capability of evolutionary algorithms can optimize the neural network parameters, ultimately leading to increased prediction accuracy.

In the combination of the MLP and the PSO algorithm, the goal of training the neural network using PSO is to find the optimal weight values for the neurons in such a way that the network error is minimized. Therefore, the training process must be optimized efficiently. In the PSO algorithm, each solution (population member) is represented as a vector of weight values. An objective function is used to evaluate each vector. The steps of the proposed method are as follows:

- Initial Preparation Phase:
  - Each population member consists of a vector containing neural network weights.
  - These weights are randomly initialized and then sent as input to the objective function.
  - The objective function receives the input vector, assigns it as the neural network weights, and applies the training inputs to the constructed network.
  - Based on the network output and the target output, the network error is calculated and returned.
- Iteration Phase:

- Each population member (weight vector) is updated using the position and velocity operators in PSO.
- The updated weights are reevaluated using the objective function.
- This process iterates until the stopping conditions are met.

- Final Selection:
  - After training, the algorithm returns the population member with the best weight values, meaning the one that minimizes the mean squared error (MSE) of the neural network.
  - The selected solution represents the optimal set of weights, ensuring improved accuracy in predictions.

The structure of the proposed method in this research is presented in a flowchart in Figure 1.

4- Results

In this section, the method for analyzing the collected data in each phase of the proposed algorithm is described. The statistical population used in this study consists of infertility data for women with thalassemia, which were collected from a thalassemia clinic in Tehran in 1403. This section focuses on analyzing the collected data from patient records based on the proposed methodology to answer the research questions. The main research questions are as follows:

1. How can a neural network be integrated with the Particle Swarm Optimization (PSO) algorithm to detect infertility?
2. What parameters have the most significant impact in detecting infertility in individuals with thalassemia?
3. How can the relationship between thalassemia and infertility be discovered?

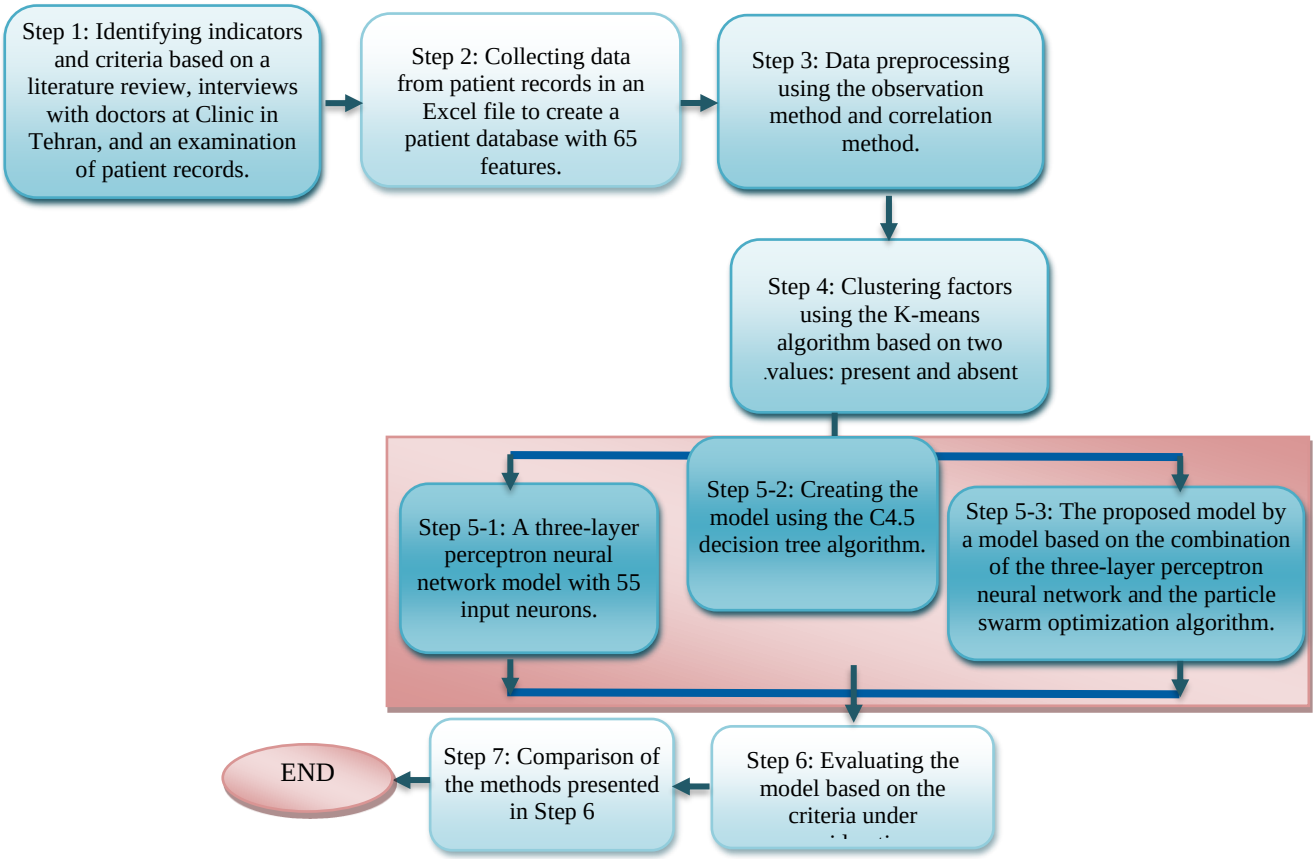

Figure 1- Flowchart of the proposed method

To evaluate the results, the following software tools were utilized: 1- Excel for organizing and preprocessing data. 2- Weka for performing data mining, feature selection, and applying machine learning algorithms. 3- Matlab for advanced modeling, data analysis, and optimizing the neural network using PSO. Patient records from a clinic in Tehran were first entered into Excel. Then, the data analysis process was carried out based on the proposed methodology. Finally, the results from these evaluations were assessed for their effectiveness in detecting infertility in individuals with thalassemia

4-1- Data Preparation and Preprocessing

This section consideres below criteria. It is appropriate for computing accuracy and addressing classification problems.

Accuracy: Accuracy measures the proportion of correctly classified instances out of the total instances. It is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where: TP (True Positive): Correctly predicted positive cases, TN (True Negative): Correctly predicted negative cases, FP (False Positive): Incorrectly predicted positive cases and FN (False Negative): Incorrectly predicted negative cases.

Precision (Positive Predictive Value): The proportion of correctly predicted positive cases out of all predicted positive cases.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall (Sensitivity): The proportion of correctly predicted positive cases out of all actual positive cases.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Root Mean Squared Error (RMSE):RMSE is commonly used in regression problems and measures the difference between actual and predicted values. It is given by:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \quad (4)$$

where: y_i is the actual value, $\hat{y}_i$ is the predicted value and N is the total number of observations.RMSE provides insight into the magnitude of prediction errors, with lower values indicating better performance.

4-2- Data Preparation and Preprocessing

This study utilizes patient data from a clinic in Tehran. According to Step One in Figure 1, relevant patient data was used. In Step Two, based on the available patient records in a clinic in Tehran, information was extracted and organized into an Excel file to create a patient database. A total of 500 patient records were examined, out of which 209 records were selected due to the completeness of all required tests. According to Step Three, to enhance performance and address incomplete information in certain features, some attributes were removed based on the observation method. These excluded attributes include:

- Personal information (e.g., first name, last name)
- Gender field (as all tests were conducted on women)
- Medical test parameters: Progesterone, Testosterone, HBA, HBF, HBS, HBH, HBA2
- Descriptions of performed procedures

Additionally, Pearson correlation analysis was applied to eliminate redundant columns with high correlation. The following correlated attributes were removed:

- Date of birth (correlated with age)
- Spouse's health status (correlated with spontaneous pregnancy)
- Weight (correlated with height)

To ensure effective training for the models, a sufficient number of records is essential. A total of 65 features were extracted from the patient records, out of which 54 features were selected based on their significance. For model evaluation, the dataset was split into: 70% training data and 30% testing data.

4-3- Clustering Criteria Using the K-Means Method

At this stage, corresponding to Step Four, the data undergo the clustering process. One of the key challenges in clustering is determining the optimal number of clusters, which in most algorithms, including K-Means, must be defined by the user. In this study, based on the volume and type of available data, we consider two clusters for comparison:

- Cluster "Have": Thalassemia patients who experience infertility.
- Cluster "Have Not": Thalassemia patients who do not experience infertility.

As shown in Table 2, clustering was performed with two clusters. This variable is designated as the 55th variable, where variables 1 to 54 serve as predictor variables, and variable 55 is the target variable.

Table 2 – Data Clustering for 2 Clusters in K-Means

Cluster 1	Cluster 2
103	106
49%	51%

4-4- Analysis of Criterion Prioritization Based on the Entropy Method

In this section, the Information Gain (IG) value is calculated for all criteria, and these values are presented in Table 2. The higher the Information Gain, the higher the priority of the criterion. All criteria are analyzed using the C4.5 decision tree, perceptron neural network, and the combination of particle swarm optimization (PSO) and perceptron neural network. The remaining criteria, whose values are not shown in Table 3, have an Information Gain of zero.

Table 3 - Information Gain Values for Criteria

Row	Feature	Information Gain	Priority of criteria	Row	Feature	Information Gain	Priority of criteria
1	Spontaneous egnancy	0.8926	1	2	Number of children	0.387	2
3	Drug infertility	0.2894	3	4	WBC (White Blood Cells count)	0.1907	4
5	Miscarriage	0.1725	5	6	Number of samples	0.1427	6
7	Beta thalassemia genetics	0.1352	7	8	XMN1 (a genetic marker)	0.1144	8
9	Cardiac MRI (cardMRI)	0.1017	9	10	Age of blood transfusion initiation	0.0863	10
11	Alpha thalassemia genetics	0.0851	11	12	CT2MS (medical test or condition)	0.0761	12
13	FBS (Fasting Blood Sugar)	0.0738	13	14	Age at marriage	0.0682	14
15	Type of thalassemia	0.066	15	16	Age at spleen removal	0.059	16
17	Type of iron chelation therapy	0.0586	17	18	Transfusion history	0.0553	18
19	Hepatitis status	0.0514	19	20	Hydroxyurea consumption	0.0466	20

4-5- Evaluation of Research Criteria Based on the C4.5 Decision Tree Method

To calculate the evaluation criteria, the data was assessed based on data validation. The results of the C4.5 decision tree evaluation for the examined criteria are presented in Table 4, and the confusion matrix for the C4.5 decision tree is shown in Table 5.

Table 4 - Comparative Prediction Values by the C4.5 Decision Tree Model

Row	Criterion Title	Criterion Value
1	Number of correctly classified samples	63 (100%)
2	Number of incorrectly classified samples	0 (0%)
3	Root Mean Square Error (RMSE)	0.0216
4	Total number of samples	63

Table 5 - Confusion Matrix of the C4.5 Decision Tree

	Cluster 1	Cluster 2
Cluster 1	32	0
Cluster 2	0	31

4-6- Evaluation of Research Criteria Based on the Perceptron Neural Network Method

Following the model construction according to Step 5-2 in Figure 1, we analyze the proposed method based on the Perceptron Neural Network. The prediction results are divided into two categories: thalassemia patients who are fertile or infertile within the specified period. The data was analyzed in a three-layer perceptron neural network, with 70% used as training data and 30% as test data. The evaluation criteria introduced in Section 4-1 were applied. The results of these criteria are presented in Table 6 and Table 7.

The perceptron neural network consists of a single hidden layer. The learning rate is 0.3, the number of training iterations is 500, and the validation threshold is 20. The backpropagation algorithm was used for training. For data validation, the batch processing method was applied, where synaptic weights are updated after processing an entire training batch. This approach minimizes overall network errors. Since the network must iterate through all data multiple times until a stopping criterion is met, batch processing is more suitable for small datasets.

Table 6 - Comparative Prediction Values by the Perceptron Neural Network Model

Row	Criterion Title	Criterion Value
1	Number of correctly classified samples	61 (96.8254%)
2	Number of incorrectly classified samples	2 (3.1746%)
3	Root Mean Square Error (RMSE)	0.0248
4	Total number of samples	63

Table 7 - Confusion Matrix of the Perceptron Neural Network Model

	Cluster 1	Cluster 2
Cluster 1	32	0
Cluster 2	2	29

4-7- Evaluation of Research Criteria Based on the Proposed Method

In the proposed method, a combination of the genetic algorithm and a three-layer perceptron neural network is used. The dataset consists of 209 samples, with 70% used as training data and 30% as test data. The population size in the particle swarm optimization algorithm is set to 100. The maximum number of generations in the genetic algorithm is 100. The crossover rate is 0.5, and the mutation rate is 0.35. The evaluation criteria introduced in Section 4-1 are used for assessment. The results of these criteria are presented in Table 8 and Table 9.

Table 8 - Comparative Prediction Values by the Proposed PSO Algorithm and MLP Model

Row	Criterion Title	Criterion Value
1	Number of correctly classified samples	63 (100%)
2	Number of incorrectly classified samples	0 (0%)
3	Root Mean Square Error (RMSE)	0.0276
4	Total number of samples	63

Table 9 - Confusion Matrix of the Proposed PSO Algorithm and MLP Model

	Cluster 1	Cluster 2
Cluster 1	32	0
Cluster 2	0	31

4-8- Comparison of the Proposed Models

Based on the conducted evaluations of the number of clusters in the K-means algorithm and the use of decision tree C4.5, multilayer perceptron (MLP) neural network, and the proposed hybrid method combining particle swarm optimization (PSO) and MLP, the models were found to be reliable in assessment. The validity and reliability of the evaluation results depend on the accuracy of the predictive models. The higher the accuracy and reliability of the models, the more valid the evaluation results.

In this study, using 70% training data and 30% testing data, the prediction accuracy was 100% for the decision tree C4.5, 96.8254% for the MLP neural network, and 100% for the PSO-MLP hybrid algorithm. As the results indicate, the combined PSO-MLP algorithm and the decision tree C4.5 provide more accurate predictions compared to the standalone MLP neural network. Therefore, it can be concluded that the hybrid PSO-MLP algorithm and decision tree C4.5 are more optimal methods compared to the MLP neural network alone. The comparison of correctly classified samples is shown in Figure 2.

Regarding the research criteria, the obtained Information Gain values indicate that out of 63 criteria, 20 have priority in the prediction process. The spontaneous pregnancy criterion ranks first and plays a crucial role in predicting thalassemia, while the number of children criterion ranks second in its influence on thalassemia prediction. The priority levels of the research criteria are illustrated in Figure 3.

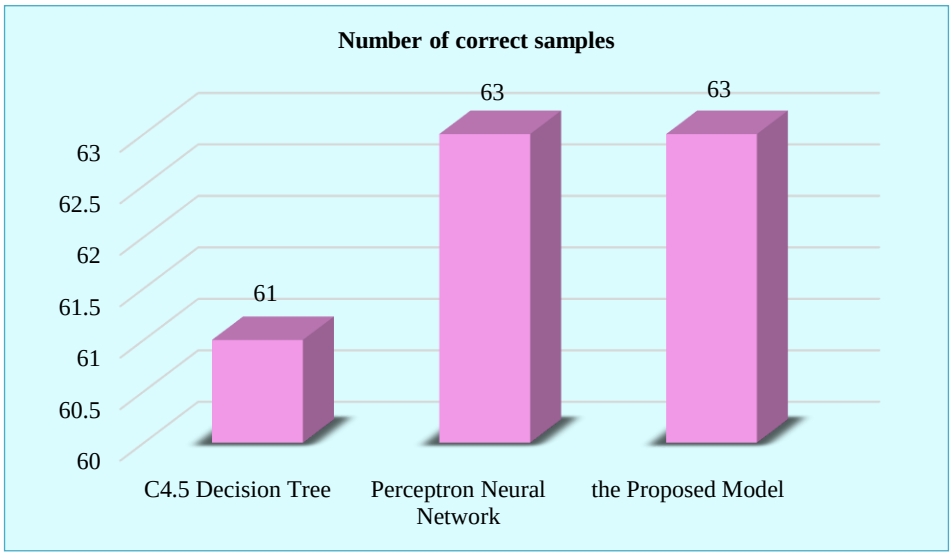

Figure 2 - Comparison of the Number of Correctly Diagnosed Samples

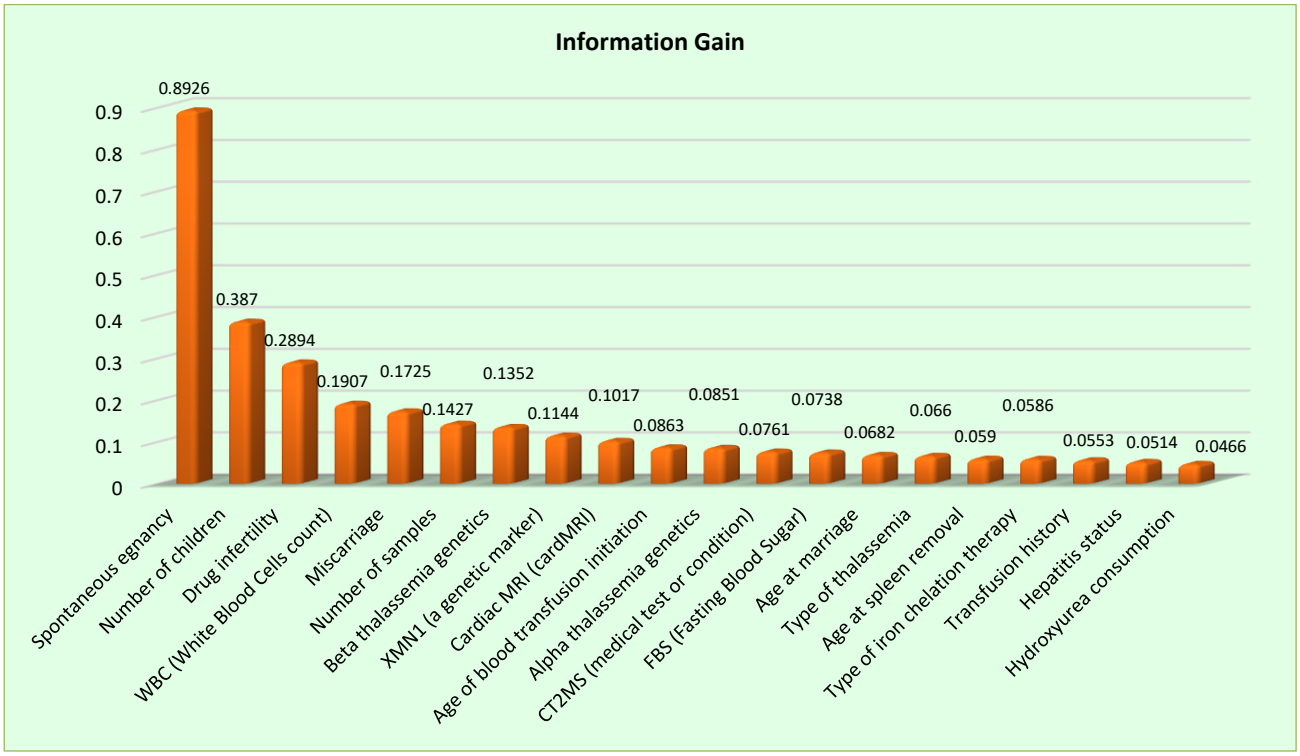

Figure 3 - Comparison of the Priority of Influential Criteria in the Prediction Process

The questions raised can be answered as follows:

1- How can a neural network be integrated with the Particle Swarm Optimization (PSO) algorithm to detect infertility?

A neural network can be integrated with the Particle Swarm Optimization (PSO) algorithm by utilizing PSO to optimize the weights of the neural network. PSO, as an optimization technique, searches for the best possible values of the weights by simulating a swarm of particles that move through the solution space, adjusting their positions based on personal and global bests. In the context of detecting infertility, the PSO algorithm can help the neural network find the optimal weights that minimize the error between the predicted outputs (infertility diagnosis) and the actual data, improving prediction accuracy.

2- What parameters have the most significant impact in detecting infertility in individuals with thalassemia?

The parameters that most significantly impact infertility detection in individuals with thalassemia include:

- Age of the individual
- Medical history of the individual (such as previous pregnancies, miscarriages, or complications)
- Hormonal levels (e.g., progesterone, testosterone)
- Hemoglobin levels (such as HbA, HbF, HbS, etc.)
- History of organ damage (such as damage to reproductive organs from thalassemia or its treatment)
- Marital status and reproductive health history of the spouse

These parameters are critical in determining whether infertility is present, as they directly influence the reproductive capabilities of individuals affected by thalassemia.

3- How can the relationship between thalassemia and infertility be discovered?

The relationship between thalassemia and infertility can be discovered through a data-driven approach. This can be done by analyzing patient records that include both thalassemia status and infertility diagnoses. Statistical analysis, such as correlation analysis, can help identify patterns and relationships between thalassemia-related parameters (e.g., hemoglobin levels, organ health) and infertility. Machine learning models (e.g., decision trees, neural networks) can also be employed to predict infertility based on thalassemia data. The most significant features influencing infertility can be identified through feature selection techniques or using optimization methods such as PSO. Through these approaches, the relationship between thalassemia and infertility can be quantified and better understood.

5- Conclusion

In this study, we examined the theoretical foundations, research methodology, and real data analysis, focusing on a proposed method based on the C4.5 decision tree, perceptron neural network, and the combination of particle swarm optimization (PSO) and perceptron neural network. Ultimately, based on the numerical results obtained through simulations, the findings were extracted. To measure infertility in thalassemia patients, a comprehensive set of criteria used for diagnosing infertility in these patients was first identified and analyzed. Based on these criteria, an evaluation

model using the decision tree, perceptron neural network, and the combination of PSO and the perceptron neural network was developed in a clinic in Tehran. The studied statistical population included patients from the year 1403 (2024–2025), whose records and documents were available at a clinic in Tehran. The infertility model for thalassemia patients was categorized into 54 criteria. Among these criteria, textual, incomplete, and highly correlated features (above 85%) were identified and removed. To prevent overfitting, which could increase network errors, the dataset was split into 70% training data and 30% testing data.

One of the challenges in decision tree modeling is determining the priority and order of features in the nodes. To address this, mathematical criteria were used. The most well-known criterion for building a decision tree is Information Gain, which is calculated by computing the Entropy of the entire training dataset and subtracting the Entropy of a specific attribute. In this study, Information Gain was computed for all criteria, and the results are presented in Table 2. The higher the Information Gain, the more important the feature.

A three-layer perceptron neural network was used, consisting of an input layer, a hidden layer, and an output layer. The hidden layer, denoted as "a," contained 63 neurons. The learning rule of this network was iterative, and the activation function was a threshold function. For modeling, the C4.5 decision tree was used. The 54 introduced variables were considered as predictor variables, and infertility (yes/no) was designated as the target variable. The dataset was divided into training and testing subsets, where the training data were used to build the model, and the test data were used to evaluate its performance.

The prediction accuracy, using 70% training data and 30% test data, was as follows: C4.5 decision tree: 100%, Perceptron neural network: 96.8254%, Genetic algorithm with perceptron neural network: 100%. As the results indicate, the combined method of the genetic algorithm and perceptron neural network, along with the C4.5 decision tree, provides more accurate predictions than the perceptron neural network alone.

Future research could explore the following: 1- Combining other evolutionary algorithms with the perceptron neural network. 2-Using multi-objective evolutionary algorithms to address infertility in women with thalassemia. 3-Applying different objective functions for various problem domains in the proposed method.

6- Refrences

- AlAgha AS, Faris H, Hammo BH, Al-Zoubi AM. Identifying β-thalassemia carriers using a data mining approach: The case of the Gaza Strip, Palestine. Artif Intell Med. 2018 Jun;88:70-83. doi: 10.1016/j.artmed.2018.04.009. Epub 2018 May 3. PMID: 29730048.
- Aessopos A, Karabatsos F, Farmakis D, Katsantoni A, Hatziliami A, Yossoef J, et al. Pregnancy in patients with well-treated Beta-thalassemia: Outcome for mothers and newborn infants. Am J Obsets Gynecol; 1999. 180(2): 360-5.
- Behram RE, Kligman RM, Jenson HB. Nelson text book of pediatrics. 18th ed. W.B Saunders Company: Philadelphia, 2008.p.2033-2037.
- El-Halees, Alaa & Alshami, Iyad. (2012). Automated Diagnosis of Thalassemia Based on DataMining Classifiers. 10.13140/RG.2.1.3336.1362.
- Ferih, K., Elsayed, B., Elshoeibi, A. M., Elsabagh, A. A., Elhadary, M., Soliman, A., Abdalgayoom, M., & Yassin, M. (2023). Applications of Artificial Intelligence in Thalassemia: A Comprehensive Review. Diagnostics, 13(9), 1551. https://doi.org/10.3390/diagnostics13091551.
- Gibbons R, DR Higgs JM, Old, Nancy F. Olivieri, Swee Lay Thein, W.G.Wood. The thalassemia syndrome, 4th edition, London D, J Weatherall and JB Clegg; 2001: 402-3.
- Kaur.G, Garg.V.K, (2025) A Comparative Review on Machine and Deep Learning Techniques for Thalassemia Identification and Classification, Journal of Information Systems Engineering and Management, Vol. 10 No. 3, PP.39-49.
- Liu, S. (2024). An Application of Machine Learning to Thalassemia Diagnosis. Journal of Computer and Communications.
- Phirom, K., Charoenkwan, P., Shoombuatong, W., Charoenkwan, P., Sirichotiyakul, S., & Tongsong, T. (2022). DeepThal: A Deep Learning-Based Framework for the Large-Scale Prediction of the α^{+}-Thalassemia Trait Using Red Blood Cell Parameters. Journal of Clinical Medicine *Vol.*11, No.21. PP.1-14.
- Saleem, M., Aslam, W., Lali, M. I. U., Rauf, H. T., & Nasr, E. A. (2023). Predicting Thalassemia Using Feature Selection Techniques: A Comparative Analysis. Diagnostics (Basel, Switzerland). Vol.13. No.22. PP.3441-3452.
- Tampakoudis P, Tsatalas C, Mamopoulos M, Tantanassis T, Christakis JI, Sinakos Z. (1997) Transfusion-depended homozygous beta thalassemia major: successful pregnancy in five cases. Eur. J. Obstet Gynecol Reprodbial. Vol.74. No.2. PP.127-3.
- Surbek D, Koller A, Pavic N. (1996) Succesful twin pregnancy in homozygous beta ovulation induction with growth hormone and gonadotropin. Fertil Steril. Vol.65. No.3. PP.670-2.