



Modifying Meta-Heuristic Algorithm to Identify the Most Important Features for Diagnosis of the Thalassemia Data in Baghdad

Fadhaa O. Sameer*

Tropical-Biological Research Unit, College of Science, University of Baghdad, Baghdad, Iraq.

Article's Information	Abstract
Received: 29.05.2025 Accepted: 28.12.2025 Published: 15.09.2026 Keywords: Features subset selection, Meta-heuristic algorithms, Hybrid machine learning model, Classification.	Machine learning feature selection is crucial for improving classification performance by breaking down high-dimensional data into manageable low-dimensional subsets. AI techniques have significantly improved the diagnosis of hematological illnesses, but little research has explored hybrid machine learning approaches that combine qualitative and quantitative data sources. This study introduces a hybrid feature selection method called ABC-PSO, which combines the strengths of artificial bee colony (ABC) and particle swarm optimization (PSO) algorithms. The method was tested against well-known meta-heuristic techniques like genetic algorithm (GA), PSO, ABC, and grey wolf optimizer (GWO) to improve classification accuracy in medical diagnosis. ABC-PSO uses both ABC's exploration potential and PSO's quick convergence to determine the most relevant subset of features. The ABC-PSO algorithm, which was tested on 300 Baghdadi thalassemia patients, demonstrated superior classification performance across various classifiers like support vector machine (SVM), random forest, and gradient boosting. In terms of accuracy, AUC, and F1-score, it fared better than standalone methods, indicating its promise as a formidable feature selection method in the analysis of medical data. Using the random forest classifier, the algorithm achieved the greatest accuracy of 0.90 and an AUC of 0.95, outperforming other algorithms consistently across all assessment criteria. The algorithm's exceptional generalization ability and durability in diagnosing thalassemia are also emphasized.
http://doi.org/10.22401/ANJS.29.3.19 *Corresponding author: fadhaa.sameer@sc.uobaghdad.edu.iq  This work is licensed under a Creative Commons Attribution 4.0 International License	

1. Introduction

Meta-heuristic algorithms have become increasingly effective in feature subset optimization in medical data analysis, especially in high-dimensional datasets where significant patterns may be obscured by duplicate or unnecessary characteristics[1]. These adaptable, population-based search techniques efficiently explore vast solution spaces without becoming stuck in local optima. Selecting the most pertinent characteristics from patient data can significantly enhance the effectiveness of machine learning classifiers in medical diagnostics, especially for complicated illnesses like thalassemia, where traditional approaches may struggle[2]. Thalassemia, a hereditary blood illness characterized by faulty hemoglobin production,

causes anemia and other associated disorders [3]. Early identification and categorization of thalassemia types (primarily major, intermediate, and minor) are critical for treatment planning and genetic counseling, especially in high-prevalence areas such as Iraq[4]. Due to the disease's complicated structure, which incorporates genetic, metabolic, and demographic components, Machine learning (ML) has become an attractive tool for diagnosis and risk assessment [5]. By picking pertinent features from noisy, challenging-to-calculate, incorrectly scaled, or correlated data attributes, the feature subset selection problem improves the effectiveness of algorithms for machine learning (ML)[6]. The objective of conventional feature selection techniques is to get rid of these

unwanted traits because machine learning is becoming more and more integrated into everyday decision-making. A major obstacle to using machine learning on biomedical datasets is the presence of high-dimensional features that may be redundant or irrelevant. This can have an impact on the interpretability and performance of the model[7]. Feature subset selection (FSS) solves this issue by determining the most useful features, improving classification accuracy, decreasing over fitting, and speeding up computation. For this problem, meta-heuristic algorithms were employed. Because of their resilience and versatility in these situations, meta-heuristic algorithms like Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Grey Wolf Optimizer (GWO), and Artificial Bee Colony (ABC) have gained widespread use. This work presents a brand-new ABC-PSO hybrid feature selection technique designed especially to increase classification accuracy in a sample of 300 Baghdad thalassemia patients. The suggested strategy seeks to maximize feature subsets that improve diagnostic performance across several machine learning classifiers by fusing the rapid convergence of PSO with the exploratory power of ABC. The rest of the paper is arranged in this order: The related study that used meta-heuristic algorithms for feature selection in the medical field in general and thalassemia in particular was illustrated in Section 1. Section 2 summarizes earlier research on feature selection using meta-heuristic algorithms for characteristics chosen in the medical field generally and in thalassemia illness specifically. Section 3 presents the materials and methods. The suggested approach is given in Section 4. One technique for choosing the subset of characteristics is the ABC-PSO hybrid algorithm. The experimental findings and discussion that demonstrated the effectiveness of the suggested approach are presented in Section 5. Section 6 concludes the article and outlines its future directions.

2. Related work

This section described the application of meta-heuristic algorithms for FSS in thalassemia data processing from 2019 to 2025. The review covers algorithmic frameworks, applications in real-world thalassemia datasets, and performance assessments. Meta-heuristic algorithms are sophisticated optimization approaches based on natural and evolutionary processes. They are frequently used in feature selection to tackle combinatorial issues, with the objective of finding the best collection of features that optimizes classification performance. The study of [8] aimed

to enhance heart disease prediction accuracy using meta heuristic optimization techniques and a machine learning model. The Cleveland heart dataset was used by [1] to measure accuracy using GAORF, PSORF, and ACORF techniques. Over the past five years, meta- heuristic algorithms like Particle Swarm Optimization, Genetic Algorithms, Grey Wolf Optimizer, and Artificial Bee Colony have significantly improved machine learning model performance. For gene selection, the paper of [9] presented the binary horse herd optimization algorithm (BHOA) in combination with the minimum redundancy maximum relevance (MRMR) filter. The study reported in reference [10] used meta-heuristic-based feature selection in machine learning for sarcopenia diagnosis, using decision trees, random forests, support vector machines, and naïve Bayes algorithms. The hybrid technique outperforms current algorithms in accuracy and minimal feature selection. For application in Thalassemia data between 2019 and 2025, various research investigated the use of meta-heuristic approaches for categorizing thalassemia types and screening for carrier status, particularly in high-prevalence areas such as South and West Asia. In study of [11] the predictive model for thalassemia risk was developed using supervised machine learning algorithms, simulated using the Waikato Environment for Knowledge Analysis (WEKA), and validated using historical data from hospitals, highlighting the importance of thalassemia parameters and risk. The study by [12] successfully tested a new method for discriminating between IDA and β -thalassemia trait using a Dynamic Harmony Search. The system uses complete blood count as input and compares it with other models like STROGANOFF, ANN, ANFIS, SVM, KNN, and traditional methods. The study [13] used the neighborhood component analysis feature selection (NCA) method to select features from datasets, achieving a 97% AUC. The RBC markers from the study performed better than those found in literature, allowing the investigation of distinct effects of each index parameter on the machine learning model and new characteristics. The study [14] used Bayesian tree-based methods to differentiate between beta-thalassemia (β TT) and intra-acute ischemic disease (IDA) in 907 Iranian patients over 18 years old. The BLTREED model and CART were used, with MCV as the main predictor. Based on CBC data, the study [3] offered a fuzzy-based model for predicting thalassemia in children that reliably identifies major, intermediate, and mild thalassemia types and compares results to medical professionals' assessments. The study

[5] introduced a Grey Wolf Optimization-based strategy for detecting blood illnesses like thalassemia, improving classification accuracy by focusing on significant characteristics. Researchers developed deep neural networks (DNNs) to improve early identification and diagnosis of thalassemia by focusing on significant characteristics, surpassing established statistical approaches [15]. The study [16] evaluated classification models using five methods and nine classifiers. The χ^2 approach achieved good accuracy with 92.65% f-score, 91.04% recall, and 91.56% precision. Combining over-sampling, synthetic minority over-sampling technique, RFE, and 10-fold cross-validation increased detection accuracy for α T patients. Hoseini et al. introduced a feature selection strategy combining support vector machines and principal component analysis, utilizing K-nearest neighbor and Naïve Bayes methods, and evaluated its performance using three datasets from the UCI Machine Learning Repository [2]. In order to choose the pertinent collection of data characteristics, the suggested approach, known as ABC-PSO, is covered in the next section.

3. Materials and Methods

3.1. Feature Selection Algorithm

Feature selection algorithms play an important part in machine learning applications [17]. Several feature selection procedures rely on meta-heuristic algorithms [18]. Meta-heuristics such as Particle Swarm Optimization (PSO), Genetic Algorithms (GA), Grey Wolf Optimizer (GWO) and Artificial Bee Colony (ABC) provide efficient search capabilities for identifying optimal feature subsets from high-dimensional datasets [1].

3.1.1. Genetic Algorithm (GA)

The Genetic algorithm (GA) is a bio-inspired optimization method that mimics the genetic and natural selection processes involved in biological evolution [19]. It is used to tackle optimization and search issues by iteratively improving a population of possible solutions across subsequent generations. GA was first developed by John Holland in the 1970s and has since gained popularity in engineering, data mining, machine learning, and feature selection challenges [20].

Algorithm Steps

- i. Set up a starting population of random chromosomes.
- ii. Evaluate each individual's fitness using a stated objective function.
- iii. To pick the best-performing chromosomes as parents for the future generation, use

methods such as a roulette wheel or tournament.

- iv. Crossover, also known as recombination, involves switching chromosomes between parents to create children as follows:
 - a. One-point crossover involves cutting at a spot and swapping portions.
 - b. Uniform crossover means swapping genes with a probability.
- v. Mutation involves randomly flipping bits in binary GA or perturbing genes to create variety.
- vi. Replace the older population with a fresh generation.
- vii. Termination: Continue until a stopping condition is fulfilled (e.g., maximum generation or adequate fitness).

3.1.2. Particle swarm optimization (PSO)

A population-based meta-heuristic algorithm called Particle Swarm Optimization (PSO) is based on the social behavior of fish schools and flocks of birds. Kennedy and Eberhart created it in 1995 and is commonly used to solve optimization issues in engineering, machine learning, and feature selection [21]. PSO refers to each potential solution as a particle, and the total set of solutions as a swarm. Particles "fly" around the solution space, directed by both their individual and the swarm's global best-known positions.

Algorithm Steps

- i. Randomly initialize particle locations and velocities.
- ii. Evaluate the fitness of each particle.
- iii. Update: Each particle's personal best position (pBest) and Global Best Position (gBest).
- iv. Update velocities and locations using the equations above.
- v. Repeat steps 2-4 until you reach convergence or the maximum number of iterations.

Mathematical Formulation

Let: $x_i(t)$: Position of particle i at time t , $v_i(t)$: Velocity of particle i at time t , $pBest_i$: Best previous position of particle i , $gBest$: Global best position among all particles, w : Inertia weight (controls exploration vs. exploitation), c_1 , c_2 : Acceleration coefficients (cognitive and social components), r_1 ,

r_2 : Random numbers in the range [0,1]. Velocity update:

$$v_i(t+1) = w \times v_i(t) + c_1 \times r_1 \times (pBest_i - x_i(t)) + c_2 \times r_2 \times (gBest - x_i(t)) \dots (1)$$

Position update:

$$x_i(t+1) = x_i(t) + v_i(t+1) \dots (2)$$

3.1.3. Grey wolf optimizer (GWO)

The population-based meta-heuristic algorithm known as Grey Wolf Optimizer (GWO) takes its cues from the natural hunting strategies and leadership style of grey wolves. The GWO has been widely used to address combinatorial and continuous optimization problems. The wolves in the population are classified into four groups depending on their fitness: Alpha (best), Beta (second-best), Delta (third-best), and Omega (rest). The Alpha directs the entire population toward the ideal solution [22]. Grey wolves encircle their victim during the hunt. The position of each wolf is updated in relation to the top three wolves (Alpha, Beta, Delta).

Mathematical Formulation

Encircling prey:

$$D = |C \times X_p(t) - X(t)| \dots (3)$$

$$X(t+1) = X_p(t) - A \times D \dots (4)$$

Coefficient vectors: $A = 2a \times r_1 - a$, $C = 2 \times r_2$

where $X_p(t)$: Position of prey (alpha, beta, or delta), $X(t)$: Current position, r_1, r_2 : Random numbers in [0, 1], a : Linearly decreases from 2 to 0 over iterations. Last location update is the average of updates toward alpha, beta, and delta:

$$X(t+1) = \frac{[(X_1 + X_2 + X_3)]}{3} \dots (5)$$

3.1.4. Artificial Bee Colony (ABC)

The Artificial Bee Colony (ABC) algorithm is a meta-heuristic optimization technique inspired by honey bee swarms' foraging behavior. Proposed by Dervis Karaboga in 2005, it is widely used in continuous optimization problems, feature selection, scheduling, and machine learning tasks [23]. The algorithm consists of four components: Food Source, Employed Bees, Onlooker Bees, and Scout Bees. Each component represents a solution vector in the search space, while the other three represent a random solution to maintain diversity.

Algorithm Steps

- i. Initialization: Randomly generate the initial food sources.
- ii. Employed Bee Phase: Each bee modifies and analyzes its existing answer.
- iii. The Onlooker Bee Phase involves selecting and improving solutions depending on fitness.
- iv. Scout Bee Phase: Replacing static food sources.
- v. Memorize the best solution found.
- vi. Repeat until the halting requirements are reached.

Food Source Update Equation

For an employed bee generating a new solution:

$$v_{ij} = x_{ij} + \phi_{ij} \times (x_{ij} - x_{kj}) \dots (6)$$

where x_{ij} current solution of bee i in dimension j ,

x_{kj} : randomly chosen neighbor, ϕ_{ij} : random number in $[-1, 1]$, v_{ij} : new candidate solution.

Probability Selection (Onlooker Bees)

Onlooker bees select food sources based on their fitness:

$$P_i = \text{fit}_i / \sum(\text{fit}_n), \text{ for } n = 1 \text{ to } N$$

where fit_n : fitness of the i -th food source, P_i : probability of selecting that source.

Scout Bee Rule

If a solution is not improved for a certain number of trials, it is replaced:

$$x_{ij} = x_{\min,j} + (\text{rand}(0,1) \times (x_{\max,j} - x_{\min,j})) \dots (7)$$

where: $x_{\min,j}$, $x_{\max,j}$: lower and upper bounds for dimension j , $\text{rand}(0,1)$: random number between 0 and 1. $\text{rand}(0,1)$ is a random integer between 0 and 1, and $x_{\min,j}$, $x_{\max,j}$ are the smaller and greater boundaries for dimension j .

3.2. Classification techniques

The categorization assignment involved anticipating a certain target using training data. In this work, five classifiers were used: Random forests, decision trees, support vector machines, k-nearest neighbors, and gradient boosting classifiers. During the exam, fitness levels were determined based on both classified and accurate mistakes.

3.2.1. Support vector machines method

For applications involving regression and classification, support vector machines (SVMs) provide a dependable and accurate classification method. They distribute all data components evenly in an n-dimensional space, with the value of features denoting a specific place. The kernel technique in SVM allows for non-linear classification in high-dimensional data. It outperforms neural network approaches and is used to categorize findings. Separating data into training and testing sets, each with labels for classes along with various features, is a common step in classification systems. SVM uses training data to generate designs [2].

3.2.2. The k-nearest neighbor method

This classification model is based on non-parametric density estimation, which assumes continuous variables rather than any data distribution. This approach efficiently saves the training dataset, including all accessible vector patterns for each period. To consider a new item x, the approach begins by calculating the distance between x and all accessible items in the training set[24].

3.2.3. The decision trees

The decision tree is a categorization model that employs an iterative gathering of the most common features at each node. It subdivides interface space into decreasing subsets. Experimentation improves decision-making categorization algorithms. Top-down construction entails detecting displacement measurements for each internal node, determining if the nodes are grouped, and grouping spatial characteristics[14]. This model is used to partition datasets for each node. The process involves determining if nodes are terminal nodes and deciding whether to divide further or create a terminal with a specific class label, while using decision trees for binary classification.

3.2.4. Random forest

The statistical machine learning approach known as random forest (RF) has been shown to be effective in regression, classification, grouping, and prediction. It requires large volumes of high-quality data for successful training, which is crucial for refining algorithms, analyzing user behavior, pattern recognition, decision-making, predictive modeling, and issue resolution. Using diverse data collection methods enhances precision, originality, speed up research, and encourages creativity[25].

3.2.5. Gradient Boosting Classifier

The Gradient Boosting Classifier is a type of ensemble training that makes a strong model for prediction by integrating the forecasts of a number of weak learners, generally decision trees. GBC builds trees one at a time, with each new tree trying to fix the mistakes made by the trees that came before it. This is different from methods like Random Forests, which build trees on their own and standard the outcomes [26].

4. Proposed algorithm for features selection

The ABC-PSO hybrid algorithm combines the exploring ability of the Artificial Bee Colony (ABC) method with the quick convergence and memory retention of Particle Swarm Optimization (PSO). This hybrid incorporates PSO's velocity and best position tracking (pBest, gBest) into the ABC framework, which improves feature selection and other optimization tasks.

Algorithm Process

step 1: Within boundaries [$xmin, xmax$], initialize N bees with random positions xi. Initialize velocities $vi=0$, personal bests $pBesti=xi$, and global best $gBest$.

step 2: Use PSO rules to update velocity, and then use that velocity to update position.

$$vij(t+1) = w \cdot vij(t) + c1 * r1(pBestij - xij(t)) + c2r2(gBestj - xij(t)) \dots (8)$$

$$xij(t+1) = xij(t) + vij(t+1) \dots (9)$$

where w decreases linearly from 0.9 to 0.4, and $c1 = c2 = 2$, population size: $N=50$. And the scout limit: limit=100 trials.

step 3: Assess fitness. If pBesti and gBest have been improved, update them accordingly.

step 4: Onlooker Bee Phase:

- i. Use a fitness-based roulette wheel to probabilistically select solutions.
- ii. To each onlooker: 50%: Generate a new solution using ABC's neighbor perturbation. 50%: Use the PSO velocity update (as described above) to improve the chosen solution.
- iii. Update pBesti and gBest to reflect improved solutions.

step 5: Phases of the Scout Bee:

- i. Solutions that surpass the trial restrictions should be abandoned.

$$x_{new,j} = gBest_j + rand \cdot (x_{max,j} - x_{min,j}) \dots (10)$$
- ii. Restart the scouts close to gBest with disturbance.
- iii. To generate new solutions, reset the trial counter using perturbation.

Pseudo-code of hybrid algorithm for feature selection (ABC-PSO)

Input:

- Number of bees (N)
- Position boundaries [xmin, xmax]
- PSO coefficients: $c1 = 2$, $c2 = 2$
- Inertia weight w (linearly decreasing from 0.9 to 0.4)
- Scout limit = 100 trials
- Max iterations = MaxIter

Initialize:

- For each bee $i = 1$ to N :
- Randomly initialize position $x_i \in [xmin, xmax]$
 - Initialize velocity $v_i = 0$
 - Set personal best $pBest_i = x_i$
 - Evaluate fitness of x_i
- Set global best $gBest = \text{best of all } pBest_i$
- Initialize trial counters = 0 for all bees

Repeat until termination (e.g., MaxIter reached):

Step 1: PSO Movement

For each bee $i = 1$ to N :

For each dimension j :

$$v_{ij}(t+1) = w \cdot v_{ij}(t) + c1 \cdot rand1 \cdot (pBest_{ij} - x_{ij}(t)) + c2 \cdot rand2 \cdot (gBest_j - x_{ij}(t))$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1)$$

- Apply boundary control to x_{ij}
- Evaluate new fitness
- If fitness improved:
 - Update $pBest_i$
 - Update $gBest$ if needed

Step 2: Onlooker Bee Phase

- Use fitness-based roulette wheel selection to choose bees

For each selected onlooker bee:

- With 50% probability:
 - Use ABC neighbor perturbation:

$$x' = x_i + \phi(x_i - x_k), \text{ where } k \neq i \text{ and } \phi \in [-1, 1]$$
- With 50% probability:
 - Apply PSO velocity and position update

as in Step 1)

- Evaluate new fitness

- Update $pBest_i$ and $gBest$ if improved

Step 3: Scout Bee Phase

For each bee i :

If trial counter > limit:

- Abandon solution
- Generate new solution:

$$x_{new,j} = gBest_j + rand \cdot (x_{max,j} - x_{min,j})$$
- Reset velocity and trial counter
- Evaluate fitness
- Update $pBest_i$ and $gBest$ if improved

Step 4: Update w (inertia weight) linearly from 0.9 to 0.4.

Particle Swarm Optimization (PSO) was set up in the suggested framework using widely used and scientifically proven hyper parameters settings:

The inertia weight, w, dropped linearly from 0.9 to 0.4. Coefficient of cognition $c1 = 2$. The social coefficient $c2 = 2$. Shi and Eberhart's strategy involves a linearly decreasing inertia weight, which balances exploration and exploitation. A high initial weight encourages global exploration in early iterations, while a lower final weight promotes convergence. The PSO literature offers a good trade-off between speed and search quality by standardizing $c1$ and $c2$ values to 2 for a balanced approach that emphasizes the particle's cognitive and collective swarm knowledge [21]. The classification models were tested using k-fold cross-validation, split into five equal groups using a 5-fold cross-validation approach. Each subset was used once as the test set, with four subsets used for training and one for testing in each round. The method provides a more accurate classifier performance estimate on unseen data, decreases volatility in performance indicators, and lessens over fitting, particularly when averaging results over data folds, particularly for small sample sizes [27].

5. Results and Discussion

To binary categorize data, four identification models were used: gradient boosting classifiers, decision trees, random forests, k-nearest neighbor, and supporting vector machines (SVM). Also five meta-heuristic methods were used to assess the feature selection quality: Artificial Bee Colony (ABC), Particle Swarm Optimization (PSO), Grey Wolf Optimizer (GWO), and Genetic Technique (GA) and the suggested technique, ABC-PSO. The recommended method is evaluated using the categorization criteria. At the Ibn-Al-Balady Hospital for Genetic Diseases in Baghdad, Iraq, 300

people with thalassemia were the subjects of a cross-sectional research. Five months of data were gathered between January 3, 2025, and June 3, 2025. Patients were only allowed to participate if they (1) were Baghdad residents, (2) had a verified diagnosis of thalassemia based on genetic and hematological tests, and (3) gave their informed permission. People with concomitant chronic conditions (such as sickle cell anemia or leukemia), insufficient medical records, or recent transfusions within the month before data collection were excluded since they could have distorted clinical measures. The study had a nearly equal gender distribution, with 153 women and 147 men in the sample, reducing sampling bias. However, it did not stratify or analyze performance by sex, potentially impacting algorithm performance or biomarker expression. Table 1 described the variables of data thalassemia inflation in Baghdad city. Table 2

showed the corresponding best features for each algorithm. Also Table 3 given the Classification Performance Metrics. In Figure1 the performance of the top classifiers (SVM, random forest, and gradient boosting) under each feature selection technique (GA, PSO, ABC, GWO, and ABC-PSO) was depicted in the ROC curves. Table 4 shows the results of the Wilcoxon signed-rank test that compares ABC-PSO with other algorithms (GA, PSO, ABC, and GWO) using five evaluation metrics: accuracy, precision, recall, F1-score, and AUC. The p-values were used to determine statistical significance. Also Figure 1 showed the ROC curves for the performance of the classifiers under each feature selection technique (GA, PSO, ABC, GWO, and ABC-PSO). Figure 2 showed the P-values of ABC-PSO to the other algorithms (GA, PSO, ABC, GWO).

Table 1 Description of variables of data

Variable	Description
ID	Unique identifier for each individual (1 to 300).
City	Location of the individual. Here, all are set to "Baghdad".
Age	Age of the individual (in years), randomly between 1 and 70.
Gender	Biological sex of the individual, randomly "Male" or "Female".
Hemoglobin	Hemoglobin level (g/dL) indicating anemia or thalassemia ($\sim 10.5 \pm 2$).
RBC	Average red blood cell count (million cells/ μ L), approximately 5.0.
MCV	Mean Corpuscular Volume (fL), indicating RBC size ($\sim 75 \pm 12$).
MCH	Mean Corpuscular Hemoglobin (pg), average hemoglobin in RBCs ($\sim 24 \pm 4$).
RDW	Red Cell Distribution Width (%), variability in RBC size ($\sim 15 \pm 2$).
Ferritin	Ferritin level (ng/mL), reflecting iron stores ($\sim 50 \pm 20$).
Target Cells	Binary flag (0 = no, 1 = yes) for presence of target cells in smear.
Family History	Binary flag (0 = no, 1 = yes) if thalassemia is present in the family.
Thalassemia	Target class (0 = No Thalassemia, 1 = Thalassemia present).

Table 2: The selected features by algorithm

Algorithm	Selected Features
GA	Hemoglobin, RBC, MCV, MCH, RDW, Platelets, Literacy
PSO	Hemoglobin, RBC, RDW, WBC, Literacy
ABC	RDW, Ferritin, Family_History, Consanguinity, WBC, Platelets, Genetic Test Result, Literacy
GWO	Hemoglobin, MCV, Target Cells, WBC, Genetic_Test_Result, Literacy
ABC_PSO	Hemoglobin, RBC, MCV, MCH, RDW, WBC, Platelets, Literacy

Table 3: The classification performance metrics by algorithms.

Algorithms of features selection	Classification method	AUC	Precision	Recall	F1	Accuracy
GA	kNN	0.83	0.8	0.85	0.82	0.82
	SVM	0.91	0.85	0.89	0.87	0.87
	Random Forest	0.93	0.86	0.90	0.88	0.88
	Gradient Boosting	0.89	0.84	0.88	0.86	0.86
	Decision Tree	0.79	0.82	0.86	0.84	0.84

PSO	kNN	0.84	0.81	0.85	0.83	0.83
	SVM	0.90	0.83	0.87	0.85	0.85
	Random Forest	0.92	0.85	0.89	0.87	0.87
	Gradient Boosting	0.88	0.83	0.87	0.85	0.85
	Decision Tree	0.78	0.79	0.85	0.82	0.82
ABC	kNN	0.82	0.80	0.84	0.81	0.81
	SVM	0.88	0.82	0.87	0.84	0.84
	Random Forest	0.90	0.85	0.88	0.86	0.86
	Gradient Boosting	0.87	0.82	0.86	0.84	0.84
	Decision Tree	0.80	0.8	0.86	0.83	0.83
GWO	kNN	0.83	0.8	0.85	0.82	0.82
	SVM	0.9	0.84	0.89	0.86	0.86
	Random Forest	0.92	0.86	0.9	0.88	0.88
	Gradient Boosting	0.88	0.83	0.87	0.85	0.85
	Decision Tree	0.8	0.81	0.87	0.85	0.84
ABC-PSO	kNN	0.85	0.83	0.86	0.84	0.84
	SVM	0.94	0.86	0.90	0.88	0.88
	Random Forest	0.95	0.88	0.91	0.90	0.90
	Gradient Boosting	0.9	0.85	0.89	0.87	0.87
	Decision Tree	0.81	0.83	0.87	0.85	0.85

Table 4: The Wilcoxon signed-rank test results comparing ABC-PSO to the other algorithms (GA, PSO, ABO, GWO) across the five evaluation metrics.

Comparison	Accuracy	Precision	Recall	F1-score	AUC	Significant
ABC-PSO vs GA	0.043	0.042	0.046	0.042	0.042	Yes for (all metrics)
ABC-PSO vs PSO	0.042	0.043	0.046	0.043	0.042	Yes for (all metrics)
ABC-PSO vs ABO	0.043	0.043	0.046	0.043	0.042	Yes for (all metrics)
ABC-PSO vs GWO	0.062	0.065	0.064	0.063	0.061	No for (all metrics)

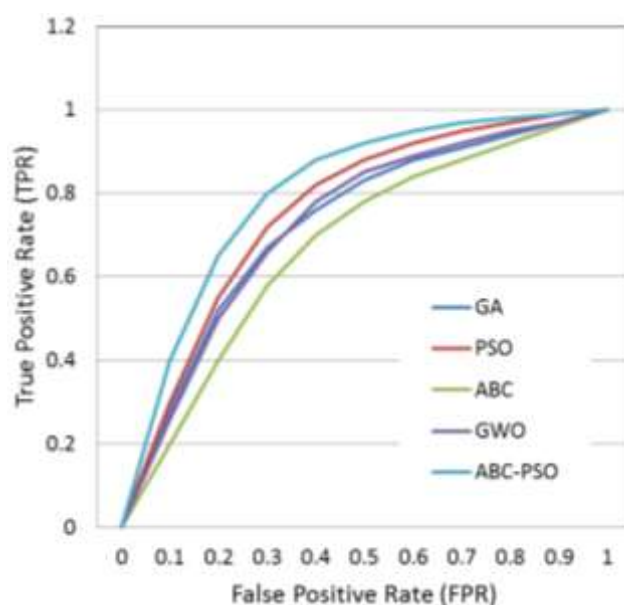


Figure 1: The ROC curves for the performance of the classifiers under each feature selection technique (GA, PSO, ABC, GWO, and ABC-PSO)

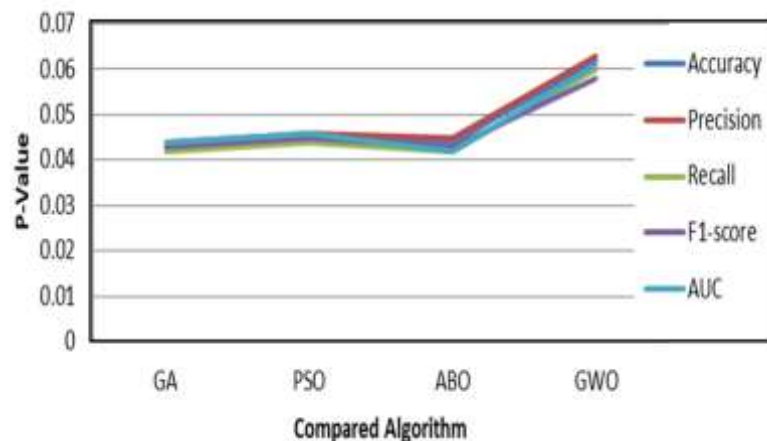


Figure 2 show the P-values of ABC-PSO to the other algorithms (GA, PSO, ABC, GWO)

Based on the examination of classifier performance across multiple optimization techniques (GA, PSO, ABO, GWO, ABC-PSO), here's a structured overview of the findings: Literacy" is frequently chosen by feature selection algorithms due to its indirect influence on illness management practices, healthcare access, and health outcomes, and its inclusion in models enhances prediction, not because it causes thalassemia. The hybrid algorithm (ABC-PSO) consistently outperforms alternative optimization algorithms in most classifiers and metrics, such as SVM, Random Forest, and Gradient Boosting. Random Forest with (ABC-PSO) scored the highest in all criteria (Accuracy: 0.90, AUC: 0.95). SVM with ABC-PSO performed best in terms of accuracy (0.88), recall (0.90), and AUC (0.94). The F1-scores and recall values exhibit a same pattern, underscoring ABC_PSO's resilience and capacity for generalization in the diagnosis of thalassemia. kNN with decision tree (ABC-PSO) has marginal but consistent benefits over alternative strategies. For kNN, the ABC-PSO has the highest accuracy (0.84), precision (0.83), F1 (0.84), and AUC (0.85). Recall is marginally greater in GA/PSO/GWO (0.85 vs. 0.86 for ABC-PSO). The GWO and GA provide excellent outcomes but fall behind ABC-PSO. In Random Forest, ABC-PSO has the greatest results (accuracy: 0.90, AUC: 0.95), showing superior class separation. Although GA and GWO perform well, ABC-PSO outperforms them. For gradient boosting, the ABC-PSO outperforms others (accuracy: 0.87, AUC: 0.90). GA and GWO produce competitive results but fall short on precision/recall. In Decision Tree, ABC-PSO and GWO have the same recall (0.87) and F1 (0.85). The ABC-PSO has the highest accuracy (0.85) and AUC (0.81), while decision trees' AUC remains low overall. AUC

Trends Decision Trees consistently underperform in AUC across all methods, implying a weaker ranking capability. ABC-PSO slightly improves this. The hybrid Advantage: ABC-PSO's combination of exploration (ABC) and exploitation (PSO) most likely explains its superior hyper-parameter tuning. The Random Forest classifier outperformed other classifiers in terms of accuracy, F1 and AUC. Random Forest is a superior classifier due to its ability to handle complex, nonlinear patterns, effectively generalize to unknown data, adjust optimal feature subsets, and ensure resilience in noisy medical settings, making it ideal for tasks like thalassemia classification due to interpretability and data complexity. Additionally, SVM and Gradient Boosting demonstrated their effectiveness in this task. Table 3 shows that ABO-PSO-RF was the best in differentiating across classes, with the greatest AUC (0.95). With an AUC of 0.93 to 0.92, GA-RF and GWO-RF likewise showed excellent performance. Across all approaches, random Forest and SVM outperformed Gradient Boosting classifiers (e.g., GA-GB). Every curve is significantly above the diagonal line, indicating that every classifier-algorithm combination outperformed the random one. From Table 4 On every criterion, ABC-PSO performs noticeably better than GA, PSO, and ABO (p-values < 0.05). Also Figure 1 shows the ROC curves for the performance of the classifiers under each feature selection technique (GA, PSO, ABC, GWO, and ABC-PSO). And Figure 2 showed the P-values of ABC-PSO to the other algorithms. The p-values for GWO are somewhat higher than 0.05, suggesting that although ABC-PSO outperforms it, the difference is not statistically significant at the 5% level. In practice, it could still have significance. All p-values for GA, PSO, and ABO are less than

0.05, indicating that ABC-PSO outperforms on every criterion. Since all of the p-values for GWO are higher than 0.05, the differences are not considered statistically significant. A dataset of 300 thalassemia patients from Baghdad, Iraq's Ibn-Al-Balady Hospital for Genetic Diseases was utilized in the study. However, generalizability could be constrained by the limited sample size and single-center design. Larger, more varied multi-center datasets should be used in future studies to increase the results' external validity.

6. Conclusions

This study shows how meta-heuristic-based feature selection can improve classification performance for diagnosing thalassemia. The hybrid ABC_PSO algorithm performs the best across a variety of classifiers and evaluation measures out of the five evaluated methods. These results highlight the promise of hybrid swarm intelligence techniques for classifying medical data, especially when dealing with high-dimensional and noisy data such as hematological characteristics. Clinically recognized important indications of thalassemia, such as hemoglobin, MCV, RDW, ferritin, and genetic test results, were among the subset of characteristics chosen by the ABC-PSO algorithm. According to medical literature, thalassemia carriers are characterized by low hemoglobin and MCV levels. Exploration rather than exploitation: The hybrid algorithm combined the global search capabilities of ABC with the convergence efficiency of PSO to avoid local optima and find a strong feature subset. ABC_PSO's statistically substantial generalization ability for diagnosing thalassemia most likely benefited from this equilibrium. This method may be used for more medical datasets and more complicated hybridizations in future studies for broader validation. Larger, more varied multi-center datasets should be used in future studies to increase the results' external validity and robustness.

Conflict of interest: The authors disclosed that they have no conflicts of interest.

Funding: No funding was received for this work.

Acknowledgement: The authors thank all medical staff of Ibn Al Balady Hospital for Genetic Diseases in Baghdad, Iraq for their help in obtaining data and providing us with information about the disease.

References

[1] Barrera, J.; Cisternas, F.; Crawford, B.; Gómez,

- S. M.; Soto, R.; "Feature Selection Problem and Metaheuristics: A Systematic Literature Review about Its Formulation, Evaluation and Applications". *Biomimetics*, 9 (1): 9, 2024.
- [2] Hoseini, S. M.; Ebtia, M.; Dehgardi, M.; "A Hybrid Feature Selection Technique Leveraging Principal Component Analysis And Support Vector Machines". *J. Artif. Intell. Data Min.*, 13 (2): 159–173, 2025.
- [3] Susanto, E. R.; Syarif, A.; Muludi, K.; Perdani, R. R. W.; Wantoro, A.; "Implementation of Fuzzy-based Model for Prediction of Thalassemia Diseases". *J. Phys. Conf. Ser.* 3rd Int. Conf. Appl. Sci. Math. Informatics Sept. 3-4, 2020, Bandar Lampung, Indones., 1751 (1): 012034, 2021.
- [4] AL-hameedawi, A. K.; Al-Shawi, A. A.; "Identification of novel mutations in β -thalassemia patients in Maysan Governorate, Iraq". *Mol. Biol. Rep.*, 50 (4): 3053–3062, 2023.
- [5] Sallam, N. M.; Saleh, A. I.; Arafat, A. H.; Abdelsalam, M. M.; "An Efficient Strategy for Blood Diseases Detection Based on Grey Wolf Optimization as Feature Selection and Machine Learning Techniques". *Appl. Sci.*, 12 (21): 10760, 2022.
- [6] Sameer, F. O.; "Modified Multi-Criteria Decision Making Methods to Assess Classification Methods". *Iraqi J. Sci.*, 64 (2): 893–906, 2023.
- [7] Sameer, F. O.; "Comparison study on the performance of the multi classifiers with hybrid optimal features selection method for medical data diagnosis". *Multimed. Tools Appl.*, 81 (13): 18073–18090, 2022.
- [8] Narasimhan, G.; Victor, A.; "A hybrid approach with metaheuristic optimization and random forest in improving heart disease prediction". *Sci. Rep.*, 15 :10971, 2025.
- [9] Mehrabi, N.; Boroujeni, S. P. H.; Pashaei, E.; "An efficient high-dimensional gene selection approach based on the Binary Horse Herd Optimization Algorithm for biologicaldata classification". *Iran J. Comput. Sci.*, 7 (2): 279–309, 2024.
- [10] Lee, J.; Yoon, Y.; Kim, J.; Kim, Y. H.; "Metaheuristic-Based Feature Selection Methods for Diagnosing Sarcopenia with Machine Learning Algorithms". *Biomimetics*, 9 (3):179, 2024.
- [11] Egejuru, N. C.; Olusanya, O. S.; Asinobi, O. A.; Adeyemi, J. O.; Adebayo, O. V.; Idowu, A. P.; "Using Data Mining Algorithms for Thalassemia Risk Prediction". *Int. J. Biomed. Sci. Eng.*, 7 (2): 33, 2019.

-
- [12] Elshaikh, R. H.; Amir, R.; Ahmeide, A.; Mohamedahmed, K. A.; Alfeel, A. H.; Higazi, H.; Hussein, S. E. O.; Babker, A. M. A.; "Evaluation of the Discrimination between Beta-Thalassemia Trait and Iron Deficiency Anemia Using Different Indexes". *Int. J. Biomed.*, 12 (3): 375–379, 2022.
- [13] Ayyildiz, H.; Tuncer, S. A.; "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". *Chemom. Intell. Lab. Syst.*, 196 (October): 103886., 2020.
- [14] Jahangiri, M.; Rahim, F.; Saki, N.; Malehi, A. S.; "Application of Bayesian Decision Tree in Hematology Research: Differential Diagnosis of β -Thalassemia Trait from Iron Deficiency Anemia". *Comput. Math. Methods Med.*, 2021 (1): 6401105, 2021.
- [15] Mo, D.; Zheng, Q.; Xiao, B.; Li, L.; "Predicting thalassemia using deep neural network based on red blood cell indices". *Clin. Chim. Acta*, 543: 117329, 2023.
- [16] Saleem, M.; Aslam, W.; Lali, M. I. U.; Rauf, H. T.; Nasr, E. A.; "Predicting Thalassemia Using Feature Selection Techniques: A Comparative Analysis". *Diagnostics*, 13 (22): 3441, 2023.
- [17] Abdulridha, S. H.; Kadhim, D. J.; "Optimal Mobile Robot Navigation in Unknown Environments using Different Optimization Techniques". *Al-Nahrain J. Eng. Sci. NJES*, 28 (2): 164–173, 2025.
- [18] Assi, J. H.; Sadiq, A. T.; "Modified artificial immune system as feature selection". *Iraqi J. Sci.*, 59 (2): 733–738, 2018.
- [19] Motair, H. M.; "Modified Firefly Algorithm using Iterated Descent Method to Solve Machine Scheduling Problems". *Al-Nahrain J. Sci.*, 26 (4): 88–94, 2023.
- [20] Samaah, F. N. U.; "A Timeless Guide to Evolutionary Problem Solving : A Review of David E . Goldberg ' s Genetic Algorithms in Search , Optimization , and Machine Learning". *J. Artif. Intell. Comput. Sci.*, 1: 16–21, 2025.
- [21] Wang, D.; Tan, D.; Liu, L.; "Particle swarm optimization algorithm: an overview". *Soft Comput.*, 22 (2): 387–408, 2018.
- [22] Mirjalili, S.; Mirjalili, S. M.; Lewis, A.; "Grey Wolf Optimizer". *Adv. Eng. Softw.*, 69: 46–61, 2014.
- [23] Nozohour-leilabady, B.; Fazelabdolabadi, B.; "On the application of artificial bee colony (ABC) algorithm for optimization of well placements in fractured reservoirs; efficiency comparison with the particle swarm optimization (PSO) methodology". *Petroleum*, 2 (1): 79–89, 2016.
- [24] Feng, X.; Xiao, Z.; Zhong, B.; Qiu, J.; Dong, Y.; "Dynamic ensemble classification for credit scoring using soft probability". *Appl. Soft Comput. J.*, 65: 139–151, 2018.
- [25] Salman, H. A.; Kalakech, A.; Steiti, A.; "Random Forest Algorithm Overview". *Babylonian J. Mach. Learn.*, 2024: 69–79, 2024.
- [26] Bowd, C.; Belghith, A.; Proudfoot, J. A.; Zangwill, L. M.; Goldbaum, M. H.; Hou, H.; Penteado, R. C.; Moghimi, S.; "Gradient Boosting Classifiers Combining Vessel Density and Tissue Thickness Measurements for Classifying Early to Moderate Glaucoma". *Am J Ophthalmol*, 217 (1): 131–139, 2021.
- [27] Gad, A. G.; "Particle Swarm Optimization Algorithm and Its Applications: A Systematic Review". *Arch. Comput. Methods Eng.*, 29 (5): 2531–2561, 2022.
-