♦ Recepción/ 27 junio 2019
♦ Aceptación/ 25 agosto 2019

Effect of Diversification and Intensification Trade-Off in Anemia Medical Data Classification

Efecto del intercambio de diversificación e intensificación en la clasificación de datos médicos de anemia

Khalid Mohammed,

College of Science, University of Diyala, Iraq

Kha2005ms@yahoo.com

Khalid Shaker

College of Computer Science and Information Technology, University of Anbar, Iraq

khalidalhity@gmail.com

Bara'a Ali Attea

College of Science, University of Baghdad, Iraq.

ABSTRACT/ Diversification intensification balancing is the major role of metaheuristic algorithms. Diversification (exploration) performs well with evolutionary metaheuristic algorithms, whereas trajectory algorithms are considered the best in local search (explanation). In this work, hybridization between differential evolution algorithm (DEA) and simulated annealing (SA) was conducted to enhance the artificial neural network (ANN) for anemia medical dataset classification. The proposed methodology begins with choosing the best ANN structure (best number of hidden layers and best number of neurons in each layer) after tackling with the hybridization of DEA and SA to obtain an improved classification accuracy. The proposed methodology registered a significant enhancement in anemia classification comparing with four other metaheuristic algorithms; the anemia data was composed of a real dataset gathered from Iraqi blood laboratories to detect anemia diseases. Meanwhile, the proposal applied two benchmarks from the University of California (UCI) repository, namely, Pima Indian diabetes data and liver disorder diseases. To verify the proposed method, we used four metaheuristic algorithms to test the selected medical benchmarks. The metaheuristic algorithms included two trajectory algorithms, namely, simulated annealing SA and Tabu search TS, and two evolutionary algorithms, namely, genetic algorithms GA and differential evolution DE. The proposed method attained remarkable results. Keywords: Metaheuristic algorithms; data mining; Medical datasets; Classification; ANN; simulated annealing; differential evolution algorithm. **RESUMEN** / **El equilibrio de intensificación de la diversificación es el papel principal de los algoritmos metaheurísticos. La diversificación (exploración) funciona bien con algoritmos metaheurísticos evolutivos, mientras que los algoritmos de trayectoria se consideran los mejores en la búsqueda local (explicación). En este trabajo, se realizó la hibridación entre el algoritmo de evolución diferencial (DEA) y el recocido simulado (SA) para mejorar la red neuronal artificial (ANN) para la clasificación del conjunto de datos médicos de anemia. La metodología propuesta comienza con la elección de la mejor estructura ANN (mejor número de capas ocultas y mejor número de neuronas en cada capa) después de abordar la hibridación de DEA y SA para obtener una precisión de clasificación mejorada. La metodología propuesta registró una mejora significativa en la clasificación de la anemia en comparación con otros cuatro algoritmos metaheurísticos; Los datos sobre la anemia se compusieron de un conju nto de datos reales recopilados de los laboratorios de sangre iraquí para detectar enfermedades de la anemia. Mientras tanto, la propuesta aplicó dos puntos de referencia del repositorio de la Universidad de California (UCI), a saber, los datos de la diabetes india Pima y las enfermedades del trastorno hepático. Para verificar el método propuesto, utilizamos cuatro algoritmos metaheurísticos para probar los puntos de referencia médicos seleccionados. Los algoritmos metaheurísticos incluyeron dos algoritmos de trayectoria, a saber, recocido simulado SA y Tabu search TS, y dos algoritmos evolutivos, a saber, algoritmos genéticos GA y evolución diferencial DE. El método propuesto logró resultados notables.**
Palabras clave: algoritmos metaheurísticos; minería de datos; Conjuntos de datos médicos; Clasificación; ANA; recocido simulado; algoritmo de evolución diferencial.

1. Introduction

Because of the ANNs ability of nonlinearity and learning, ANNs are established at a lot of medical classifications problems. In this paper artificial neural network (ANN) was hybridized with a hybridization between differential

evolution algorithm (DEA) and simulated annealing (SA) to improve anemia disease classification. DEA and SA are metaheuristic algorithms; DEA is considered as an evolutionary population-based solution algorithm, whereas SA is a trajectory method operating on a single solution. The balance between diversification and intensification is the major role of metaheuristic algorithms, according to Yang et al. [1], Blum and Roli [2], and Talib [3].

Diversification (exploration) performs well with evolutionary metaheuristic algorithms, whereas trajectory algorithms are considered the best in local search (explanation). In this work, DEA and SA were hybridized to enhance the ANN for anemia medical dataset classification. The proposed methodology begins with choosing the best ANN structure (best number of hidden layers and best number of neurons in each layer) and then enhancing the classification process by hybridizing DEA and SA to improve the classification accuracy. The proposed methodology registered a significant enhancement in anemia classification. The anemia data is a real dataset gathered from Iraqi blood laboratories to detect anemia diseases. In addition, the proposal was applied into two benchmarks from the University of California (UCI) repository, Pima Indian diabetes (PID) data, and liver disorder (LD) diseases. The proposal method registered remarkable results.

Many different data-mining algorithms in literature are used to classify several types of diseases, such as anemia disease, into specific types on the basis of the data-mining algorithms by Elshami and Alhalees [4]. A person with anemia is probably unaware of the problem because symptoms may not appear. Millions of people may suffer from anemia and their health exposed to risk. Therefore, the disease is significant; several studies carried out in this domain are mentioned in the study of Yilmaz et al. [5].

Sanap et al. [6] developed a system by using the classification technique C4.5 decision tree algorithm and Sequential minimal optimization (SMO) support vector machine using WEKA. The scholars implemented a number of experiments that uses these algorithms. The anemia classification applies a decision tree that provides clear results depending on Complete Blood Count (CBC) reports. Amin et al. [7] have compared between naive Bayes, J48 classifier, and neural network classification algorithms through WEKA and studied hematological data to specify the most appropriate algorithm.

No rule or method guarantees the balance between exploration (diversification) and explanation (intensification) in metaheuristic algorithms for all problems, as stated by Yang et al. [1]. However, different studies attempt to achieve a balance between diversification and intensification. Fagan and Vuuren [8] declared six general views of diversification and intensification terminology from literature reviews.

Another study by Makas and Yumusak [9] combined the artificial bee colony (ABC) with migrating birds optimization (MBO) to attain balance between exploration and explanation through the exploration property of ABC and explanation property of MBO under a sequential execution strategy.

2. Materials and Methods

This section will pass through materials and most used methods in the article as follow:

2.1 Artificial neural network ANN

Through mapping input data to the approximate desired output, ANN can be adopted as a classification model. This model includes an input layer (the layer that receives inputs), an output layer (the layer that provides outputs), and hidden layer(s) between them.

The attributes from disease data sets are input to ANN in this study. These inputs are examined in the input layer and multiplied by weights. The weights are randomly initialized relative to the neurons in the hidden layer(s), where the summation is specified in the activation function, as indicated in Eq. (1) and Eq. (2).

$$s(x) = \sum_{i=1}^{n} x_i w_i \qquad (1)$$

The neuron output was determined after the obtained summation results from the activation function were evaluated. The following sigmoid function was adopted in the proposed model:

$$f = \frac{1}{1 + e^{-s(x)}} \qquad (2)$$

In order to obtain reliable estimate of classifier accuracy holdout and random sampling method used to assessing accuracy. According

that the data sets in this study are divided into (40%) training, (30%) validation, and (30%) test data sets. The trained neural network structure is used to evaluate populations, and the training set of weights (w1, …, wn) are input into the metaheuristic algorithms[16][17][18].

2.2 Differential evolutionary DE

The DE initializes the random population of d-dimensional vectors. The representation of the solution applied to the DE is the same as that applied in the GA.

The DE combines several solutions with the candidate solution to generate a new solution, as explained by Storn and Price [10][19][20]

Three main operations were considered for population solution evolution through repeated cycle; these operations were mutation, crossover, and selection. Despite the similarity in naming GA operations, the operations were not exactly the same.[21]

The process of generation in every iteration followed those of Storn and Price [10]; Kachitvichyanukul [11]:

Three vectors were selected randomly from the population (not the target vector) and combined to generate mutant vector V as the first step. The combination process was performed in accordance with Eq. (Error! **No text of specified style in document.**3).

$$V = X_1 + f(X_2 - X_3) \quad \text{(Error! No text of specified style in document.3)}$$

where X_1, X_2, and X_3 are the randomly nominated vectors from the population, and F is a constant factor controlling the amplification of the differential variation ($X_2 - X_3$) and considered as the main parameter of DE.

The second step is the crossover between mutant vector and target vector. The crossover method in DE was used for either binomial crossovers in Eq. (4) to produce trial vector according specified crossover probability CR.

$$U_i = \begin{cases} V_i, & random(i) \le CR \\ X_i, & random(i) > CR \end{cases} \quad (4)$$

where U is the trial vector, V is the mutant vector, and X is the target vector. CR is the crossover rate, whereas i: 1 ... D, where D is considered as the number of dimensionalities. Figure 1 show crossover process for 8-dimension parameters.

The third step involves the selection operation where the best vector is chosen between the target vector and trial vector depending on fitness. The best fitness is targeted in the next generation.

DE functionality steps are illustrated in Figure 2, where the figure represent flowchart for the algorithmic steps of the approach.

In this study, each individual represents a set of weights for the ANN model, and fitness is reflected by the ANN accuracy.

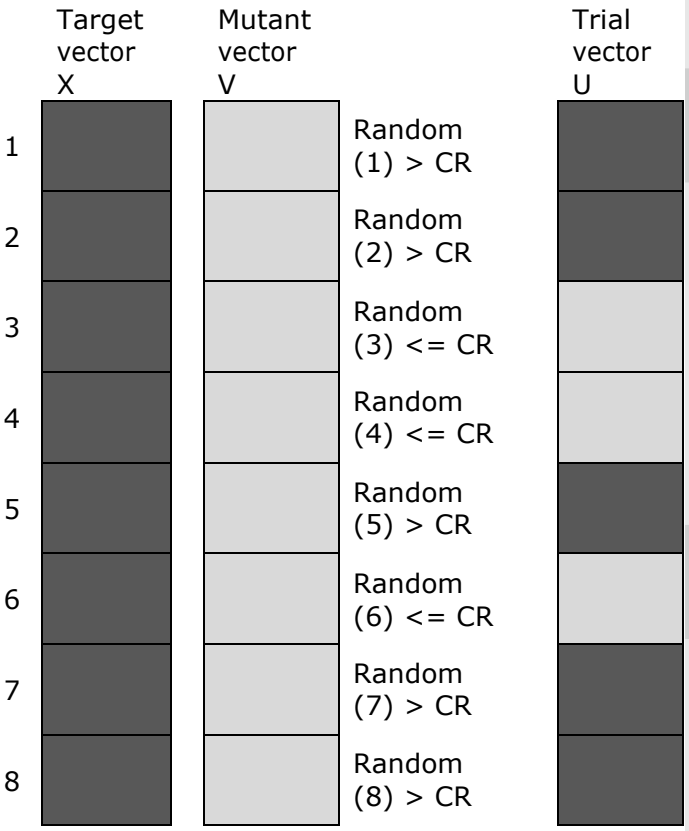

Figure 1**:** Crossover process for 8 parameters

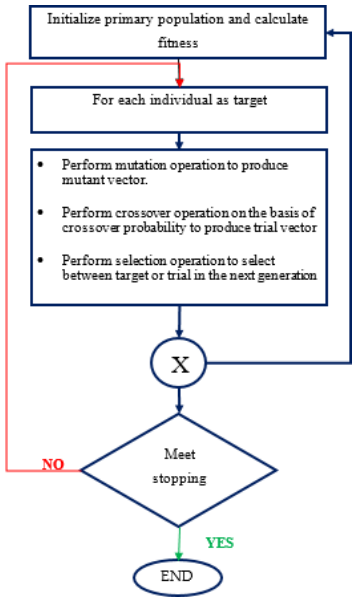

Figure 2: illustrate operation steps of differential evolution algorithm as flowchart.

2.3 Simulated annealing algorithm SA

SA functionality uses a single agent or solution that moves over the design space or search space in a piecewise style, as described by Kirkpatrick et al. [12]. Basically, SA is based on a previous method called "Metropolis algorithm" by Metropolis et al. [13]; Blum and Roli [2].

In the latter algorithm, some jobs that do not lower the range are accepted when they serve to allow the solver to "explore" further the possible space of solutions. Such "bad" trades are allowed under the criterion in Eq.(Error! **No text of specified style in document.**5):

$$e^{-\frac{}{T}} > Rand(0,1) \quad \text{(Error! No text of specified style in document.5)}$$

where D= new solution – current solution, and T: temperature.

The general form for trajectory methods can be formulated as follows:

Initialize single solution

Loop

Improve

Until stop criteria

In this study, SA's steps assist to jump local minima solutions and enforce the algorithm to explore global optimum solution as in figure 3. The SA algorithm adopted the steps illustrated in the flowchart in figure 4, as shown in the flowchart the search will depend on the value of $e^{\frac{-new\_sol}{Temp}}$, which decreases over time to provide diversification for the solution through algorithm loop iteration.

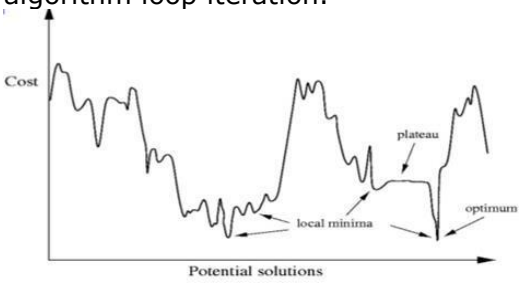

Figure 3: exploration strategy to jump local minima

3. Proposed Work

This study presents the details of hybridization between differential evolution (DE) and SA to enhance the ANN and proposes the ANN_DESA method to balance exploration (diversification) with exploitation (intensification). The proposed methodology begins with choosing the best ANN structure, best of hidden layers number, and best number of neurons in each layer. The selection is performed by testing every ANN structure on the dataset with boundaries of 30 hidden layers and 10 neurons to find the best structure with the best accuracy. Weights of the best structure (W1 … Wn) were introduced to hybridize DESA, where DE adopted an exploration role in the method, whereas SA maintained an intensification role for exploitation in the promising solutions of the search space as details in Figure 5. Resulting weights (W^1 … W^n) were tested on dataset selection for enhanced accuracy in terms of diversification solutions. The proposed study is illustrated in **Error! Reference source not found.**.

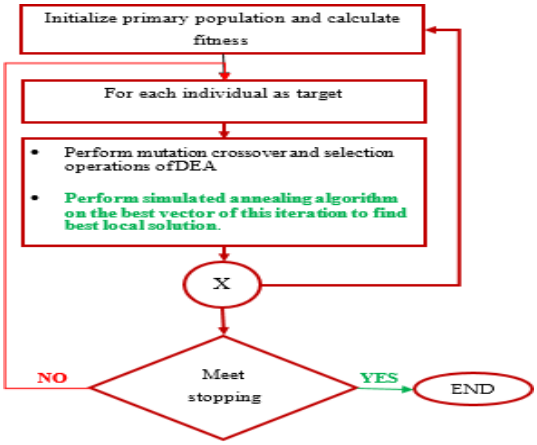

Figure 5. Hybridization between DEA and SA.

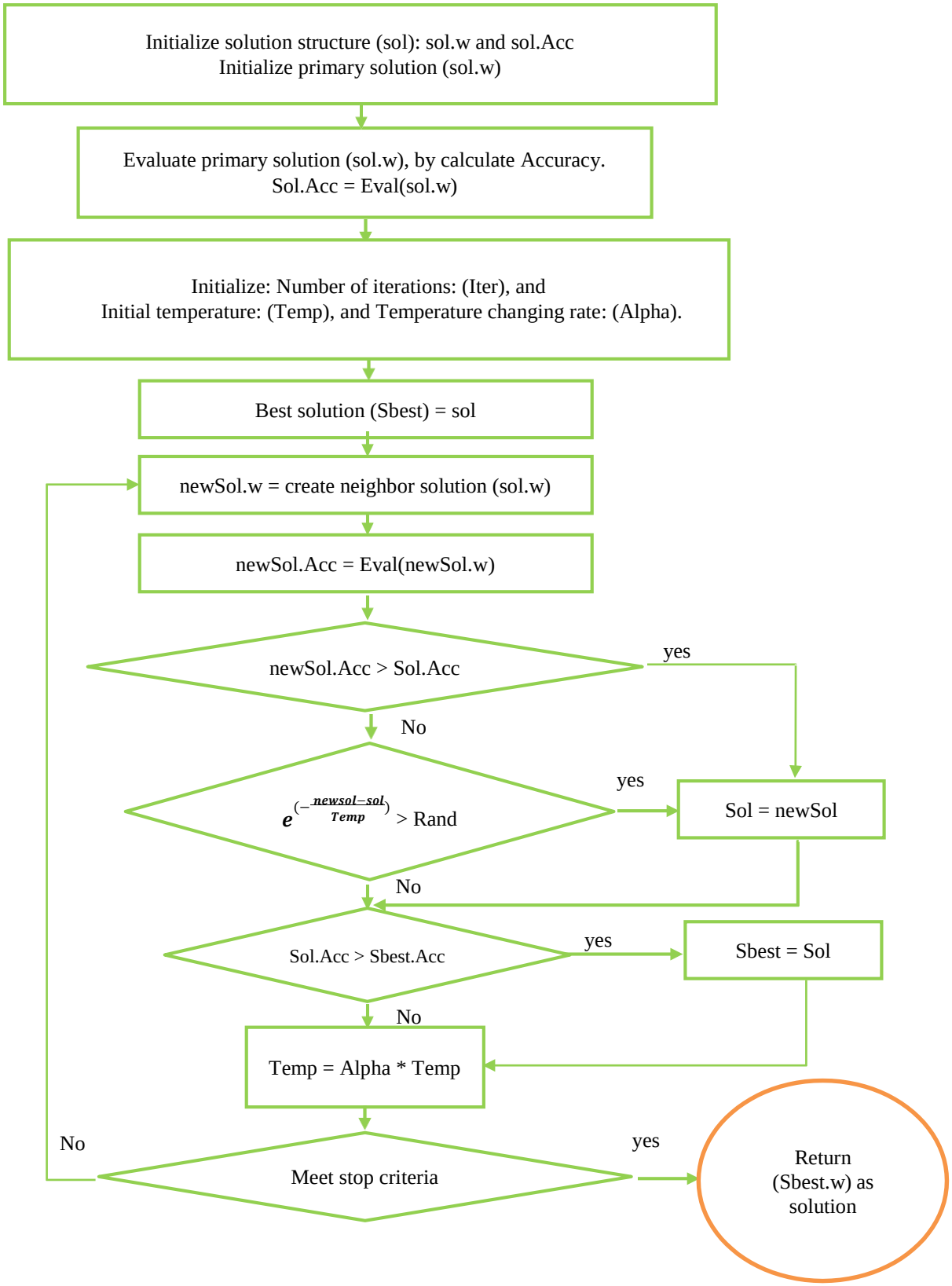

Figure 4. Simulated annealing algorithm

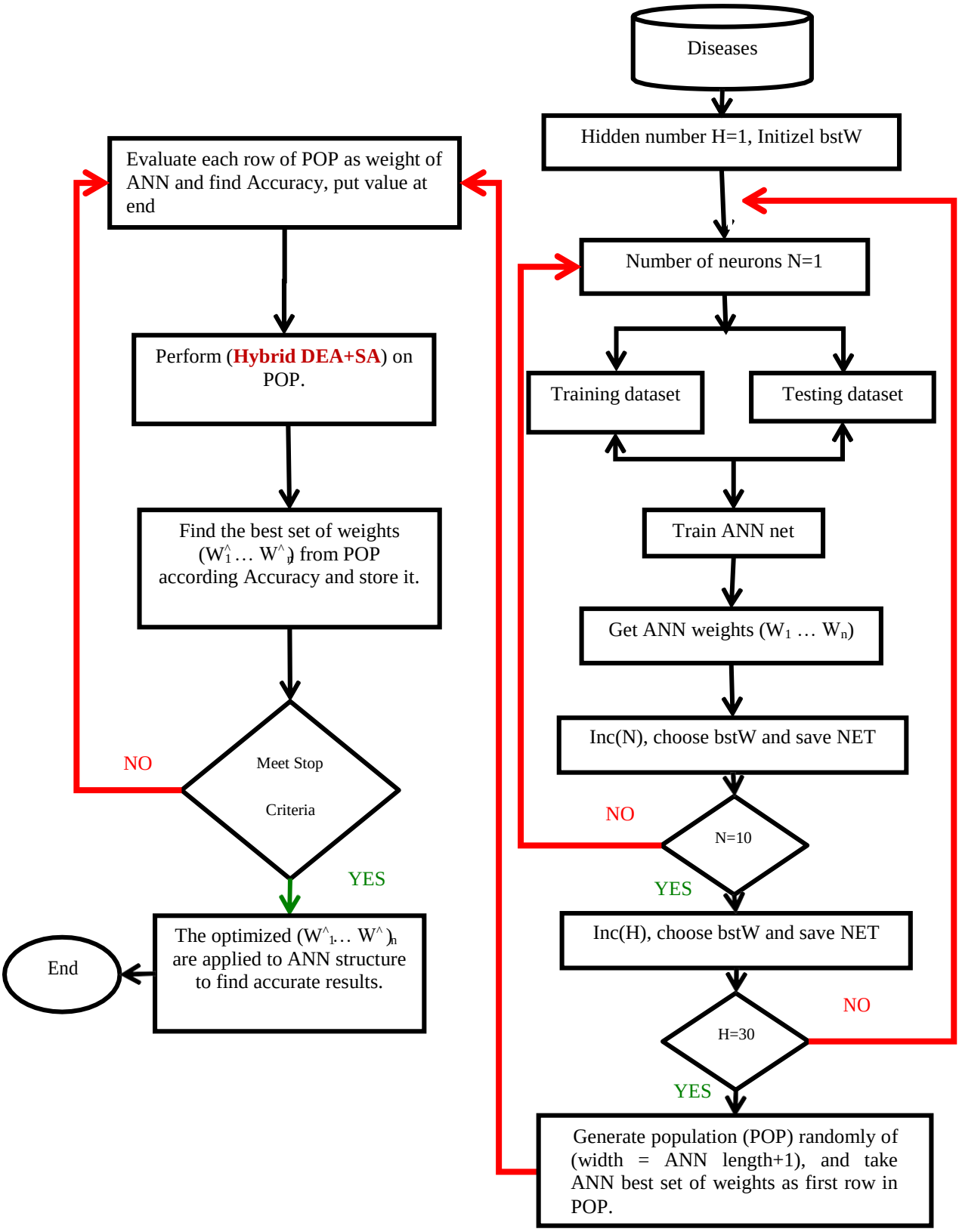

Figure 6: Proposal methodology

In order to verify the proposal method, the selected medical datasets tested on four different metaheuristic algorithms beside the proposal, two of them trajectory metaheuristic that are simulated annealing SA, and Tabu search TS, whereas the other two are evolutionary metaheuristic algorithms which are genetic algorithm GA, and differential evolutionary DE.

The parameter settings are illustrated in Table Error! **No text of specified style in document.**1, Setting for the proposed algorithms TS, SA, GA, DE, and DESA when tested on anemia, PID, and LD.

Table **Error! No text of specified style in document.**1: Parameter setting for the proposed work

DE		**GA**	
Parameters	Value	Parameters	Value
Number of generations	1000	Number of iterations	1000
size of population	50	size of population	50
Rate of crossover	0.8	Rate of crossover	0.7
		Mutation rate	0.3
TS		**SA**	
Parameters	Value	Parameters	Value
Number of iterations	1000	Number of iterations	1000
		Temp	0.025
		Alpha	0.99
DESA method setting			
DEA part parameter setting		SA part parameters setting	
Parameters	Value	Parameters	Value
Number of generations	1000	Number of iterations	100
size of population	50	Temp	0.025
Rate of crossover	0.8	Alpha	0.99

To effectively evaluate the proposed work ANN_DESA, we tested the three datasets on the ANN hybridized with two trajectory methods SA and TS and with two evolutionary algorithms GA and DE. Finally, we optimized the ANN by hybridization of DE and SA with DESA. Experimental results based on classification accuracy were obtained to measure the performance of the diverse classifiers relative to our approach. The accuracy was calculated as described in Eq. (6).

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

where TP: true positive, TN: true negative, FP: false positive, and FN: false negative.

3.1 Diseases datasets

ANN classification and optimization algorithms are hybridized on the following disease data sets:

(1) Pima Indian diabetes PID

This data set comprises patients who are pregnant females at least 21 years old and of Pima Indian heritage according to the National Institute of Diabetes and Digestive and Kidney Diseases.

Number of tuples: 768

Number of attributes: 9 (including class).

(2) Liver disorder LD

Seven attributes are included in this data set. Blood test results related to liver disorders due to alcohol consumption are indicated in the five attributes. The number of drinks per day constitutes the sixth attribute. Patient class and condition (i.e., whether the patient has the disorder or not) are presented in the seventh attribute.

Number of tuples: 345

Number of attributes: 7 (including class attribute)

(3) Anemia

real data set taken from blood laboratory in AL-Anbar health directorate / Iraq. Anemia is an indication of a low level of hemoglobin, which will cause a decrease in the level of oxygen transfer to the tissues of the human body Hoque et al. [14]. A Complete Blood Cell test conducted for patients in laboratory. Anemia data is real dataset gathered from Iraqi blood laboratories to detect Anemia diseases. The anemia diagnosing identified using this information: age, gender, hemoglobin (HP), Hematocrit (HCT) and other attribute values shown in

Table 2.

Table 2: Anemia dataset specifications (attributes)

Abbreviations	Explanation	Units
Age	Patient age	Years
Gender	Patient gender	1: male 2: female
HP	Hemoglobin	(G/dl × 10^6)
RBC	Red blood cell	(Cell/µl × 10^3)
HCT	Hematocrit	(%)
MCV	Mean corpuscular volume	(fL)
MCH	Mean corpuscular hemoglobin	(pg)
MCHC	Mean corpuscular hemoglobin concentration	(G/dl × 10^6)
WBC	White blood cell	(Cell/µl × 10^3)
PLT	Platelets	(Cell/µl × 10^3)

In general, Table shows a brief description for the datasets characteristics used in this study.

Table 3: Description of datasets

Dataset	No. of attributes	No. of records
Anemia	11	803
PID	8	768
LDs	9	286

4. The results

In the context of the study, three disease datasets, namely, PID and LD from the UCI repository and anemia dataset from the AL-Anbar health provenance laboratories /Iraq, were tested with the two trajectory algorithms SA and TS, two evolutionary algorithms GA and DE, and finally with the proposed method ANN_DESA. Empirical results are shown in Table Error**! No text of specified style in document.**.

Table **Error! No text of specified style in document.**: Accuracy of the three diseases datasets tested by ANN and the five algorithms

Medical dataset	ANN	TS	SA	GA	DE	DESA
Anemia	94.14	97.01	96.77	96.64	96.52	**97.1**
PID	78.91	80.86	80.86	82.03	80.86	**84**
LDs	75.65	78.84	78.26	77.68	78.26	**79.8**

Obviously, the results of DESA were superior to those of all the other algorithms. Hence, we concluded that the balance between the exploration and the exploitation can widely benefit the exploration of search space and intensive exploitation of the most promising solutions within the local area of the search and balance between the two operations.

Statistical t-test was performed after 25 runs of ANN against ANN+DESA to test the statistical significance of differences. The t-test statistics were transformed into a conditional probability called p-value. The p-value was less than 0.0001, and the difference was extremely significant. The test values are listed in Table .

Table 5: Statistical test of the three medical datasets PID, LD, and anemia between ANN and ANN + DESA

	PIMA	LD	Anemia
Mean difference	0.03684000	0.002676	0.010216
T score	14.1156	36.5694	11.8647
standard error of difference	0.005	0.002	0.001
Difference verdict	Significant	Significant	Significant

5. Discussion

The hidden layers number and number of neurons in every layer are critical in ANN

leaning. A suitable ANN structure was built by selecting the best number of hidden layers and the best neurons number in each layer to obtain a structure that produces the best accuracy in medical dataset classification. Boosting the ANN model complexity by raising the number of hidden layers and number of neurons may lead to overfitting. Metaheuristic algorithms normally solve the ANN overfitting problem. However, in the context of this study, ANN exclusively pushes for a potential high overfitting in the training phase. Therefore, ANN requires a metaheuristic method that can explore the search space to fulfill the diversification and simultaneously

Table **Error! No text of specified style in document.**, and Figure . This hybridization combined the two algorithms perfectly and performed more effectively than all the other algorithms.

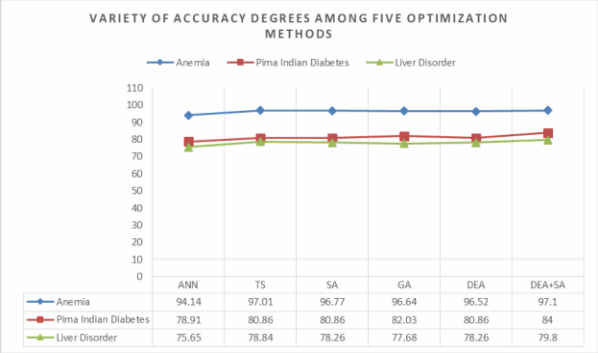

Figure 7: Accuracy degrees of anemia, PID, and LDs under testing with optimization methods

All the metaheuristic algorithms chosen in this study recorded an enhancement when hybridized with the ANN despite a low improvement. The difference in enhancement was not large but the accuracy results of the classification were high, considering literature and the study of ANN structure impact on reaching better accuracy in Salman et al. [15], This result was achieved because we have already relied on the neural network model to find the greatest value for accuracy.

6. Conclusion

This study aimed to propose a metaheuristic method to facilitate balance between exploration and exploitation and consequently enhance medical dataset classification. The number of hidden layers with the neurons number in each layer can affect ANN learning. Preparing an ANN structure by selecting a complex structure to achieve high accuracy can benefit metaheuristic algorithm efficiency perform intensive exploitation for the most promising solutions in the local areas.

The DESA method represents an example of a balance between the exploration and the exploitation. The exploration feature of the DE and the exploitation feature of the SA algorithm were combined to reach the global optimum solution.

When hybridizing two metaheuristic algorithms among a list of metaheuristic algorithms, selecting two high-accuracy algorithms is not mandatory. In anemia as an example, we noted the accuracy of TS to be greater than that of SA and the accuracy of GA to be greater than that of DE despite that best hybridization achieved between DE and SA for searching the global optimum. When combining two metaheuristic algorithms to formulate a new improved algorithm for any reason, choosing the algorithm with highly accurate result between the two is not mandatory. Instead, empirical tests can be performed to find convenience. Our approach can be adopted in future studies to solve various problems.

References

[1] X. Yang, S. Deb, and S. Fong, "Metaheuristic Algorithms : Optimal Balance of Intensification and Diversification" no. September 2014. doi:10.12785/amis/080306.

[2] C. Blum and A. Roli, "Metaheuristics in combinatorial optimization: overview and conceptual comparison," *ACM Comput. Surv.*, vol. 35, no. 3, pp. 189–213, 2003.

[3] Talbi, El-Ghazali. Metaheuristics: from design to implementation. Vol. 74. John Wiley & Sons, 2009.

[4] E. H. Elshami and A. M. . Alhalees, "Automated Diagnosis of Thalassemia Based on DataMining Classifiers," no. June 2012.

[5] Yilmaz, Ahmet, Mehmet Dagli, and Novruz Allahverdi. 2013. "A Fuzzy Expert System Design for Iron Deficiency Anemia." In *AICT 2013 - 7th International Conference on Application of Information and Communication Technologies, Conference Proceedings*. doi:10.1109/ICAICT.2013.6722707.

[6] S. A. Sanap, M. Nagori, and V. Kshirsagar, "Classification of Anemia Using Data Mining Techniques BT - Swarm, Evolutionary, and Memetic Computing," 2011, pp. 113–121.

[7] N. Amin and A. Habib, "Comparison of Different Classification Techniques Using WEKA for Hematological Data," no. 3, pp. 55–61, 2015.
[8] F. Fagan and J. H. Van Vuuren, "A unification of the prevalent views on exploitation , exploration , intensification and diversification," vol. 2, no. 3, pp. 294–327, 2013.
[9] H. Makas and N. Yumus, "Balancing exploration and exploitation by using sequential execution cooperation between artificial bee colony and migrating birds optimization algorithms," pp. 4935–4956, 2016.
[10] R. Storn and K. PRICE, "Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces," pp. 341–359, 1997.
[11] V. Kachitvichyanukul, "Comparison of Three Evolutionary Algorithms :," vol. 11, no. 3, pp. 215–223, 2012.
[12] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, "Optimization by Simulated Annealing," vol. 220, no. 4598, pp. 671–680, 1983.
[13] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of State Calculations by Fast Computing Machines," vol. 1087, 1953.
[14] M. Hoque M and M. Kader SB, "Risk factors for anaemia in pregnancy in rural KwaZulu-Natal , South Africa : Implication for health education and health promotion," vol. 51, no. 1, pp. 68–72, 2009.
[15] I. Salman, O. N. Ucan, O. Bayat, and K. Shaker, "Impact of metaheuristic iteration on artificial neural network structure in medical data," *Processes*, vol. 6, no. 5, 2018.