

Diagnosis of Coronary Artery Disease Using Artificial Bee Colony and K-Nearest Neighbor Algorithms

Ismail Babaoğlu, Mustafa Servet Kiran, Erkan Ülker, and Mesut Gündüz

Abstract—Artificial bee colony (ABC) is one of the swarm intelligence optimization algorithms, inspired by foraging and dance behaviors of real honey bee colonies. This study is an instance of a hybrid algorithm using ABC together with k-nearest neighbor algorithm on diagnosis of coronary artery disease employing exercise stress test data. The study dataset is composed of 134 healthy and 346 unhealthy totally 480 patients. On the proposed algorithm two centroid vectors are obtained concerning one for healthy patients and the other for unhealthy patients utilizing ABC for the training part of the dataset. Then, the test part of the dataset is classified using k-nearest neighbor algorithm. The results obtained by the proposed technique show that this hybrid algorithm could be used as an alternative classifier on diagnosis of coronary artery disease employing exercise stress test data.

Index Terms—Coronary artery disease, exercise stress testing, artificial bee colony

I. INTRODUCTION

In last decades, researchers develop many algorithms by inspiring stability and success of the nature for solving real application problems. These developments improved many problems solution techniques so that the success rates.

Artificial intelligence and machine learning techniques are used as classification or prediction methods in determining heart diseases primarily arrhythmia, coronary artery disease and myocardial infarction with a growing phenomena. Coronary artery disease (CAD) is the most common heart disease seen among these diseases. It is the primary cause of mortality and morbidity on developed or developing countries [1].

Coronary angiography (CAG) is the gold standard method used on diagnosis of the suspected CAD. However, it is an expensive and invasive method having a mortality rate of 2-3%. Because of these reasons it is not the first choice diagnosis method on the diagnosis of CAD. Exercise stress testing (EST) is one of the most commonly used method on the diagnosis of CAD. This method is reproducible, non-invasive, relatively cheap and safe. This is why the cardiologists commonly prefer EST instead of CAG. But EST has a relatively low sensitivity and specificity on the diagnosis of CAD, and this limits EST's clinical usage [2], [3].

Considering the use of artificial intelligence and machine learning approach on diagnosis of heart diseases, many researches can be found in literature. Maglogiannis et al. made a research that had been conducted concerning the

exploitation of heart sound signals for automated and semi-automated detection of pathological heart conditions and so the diagnosis of heart valve diseases using support vector machines (SVM) [5]. Choi had also studied on detection of valvular heart diseases using wavelet packet decomposition and support vector machine [5]. Zhidong proposed non-invasive diagnosis method of coronary artery disease based on the instantaneous frequency estimation of diastolic murmurs and SVM classifier [6]. Kukar et al. examined EST and myocardial scintigraphy data as inputs and CAG data as reference for ischemic heart disease using some common machine learning techniques [7]. Lapuerta et al. investigated artificial neural network (ANN) performance to predict the occurrence of CAD based on information from serum lipid profile [8].

In this study, depending artificial bee colony (ABC) algorithm's high performance and accuracy in a large set of optimization problems [9], ABC algorithm is integrated with k-Nearest neighbor in order to investigate classification accuracy, performance and effectiveness of the proposed method on determining CAD using EST data.

II. MATERIALS AND METHODS

A. Data Collection

Four hundred and eighty patients who underwent EST and coronary angiography (CAG) were included in the study. Baseline demographic characteristics, rest and peak exercise heart rate, blood pressure, and exercise time were recorded. The EST results were evaluated by 2 experienced cardiologists (human-based CAD assessment). ST segment depression and elevation occurred 60 ms after the J point were recorded at each derivation in peak exercise. According to human-based method, an exercise test result was considered positive if there was ≥ 1 mm horizontal or downsloping ST depression or ST elevation in two contiguous leads. Within the first month following the EST, CAG was performed to all patients, and the angiographic images were evaluated by 2 experienced cardiologists. Presence of $\geq 50\%$ narrowing in left main coronary artery, or $\geq 70\%$ narrowing in other major epicardial coronary arteries indicated significant CAD. The patients with bundle branch blocks (right or left), pre-excitation syndromes, atrial fibrillation, left ventricular hypertrophy and those taking the digoxin were excluded from the study. The EST dataset was obtained from Selçuk University Meram Faculty of Medicine and the descriptive features of the dataset are ST state changes and heart rate at resting, age, sex, peak stage (1 to 5), exercise duration (in seconds), heart rate at peak stage, metabolic equivalent, reason for termination (target heart rate,

ST depression or others), chest pain, previous myocardial infarction and values of the 12 derivations (D1, D2, D3, aVR, aVL, aVF, V1, V2, V3, V4, V5, V6) at peak stage which obtained from the EST, and presence of coronary artery disease which obtained from CAG.

B. Artificial Bee Colony

Artificial bee colony algorithm is one of the commonly used swarm intelligence algorithm proposed by Karaboga for solving continuous optimization problems [10,11]. ABC algorithm simulates intelligent behaviour of real honey bee colonies. There are three kinds of bees named as *employed bees*, *onlooker bees* and *scout bee* in the ABC hive. Employed bees try to improve self-food source which is feasible solution for the optimization problem and generated in initialization of the algorithm. The employed bees forage food sources and move position information about them to the hive. The onlooker bees search around the solutions of employed bees by considering information shared by employed bees. The third type of bees is scout bee. If an employed bee could not improve self-solution in a certain time name as limit which is determined for the population in initialization of the algorithm, this employed bee becomes a scout bee. After a new solution is produced for this scout bee, the scout bee becomes employed bee, again.

In order to update the food source positions, the employed bees uses equation given as follow:

(Let $V_i = X_i$)

$$V_{i,j} = X_{i,j} + \varphi(X_{i,j} - X_{k,j}) \quad (1)$$

$$i \neq k \text{ and } i, k \in \{1, 2, \dots, N\}$$

where, $V_{i,j}$ is the candidate food source position for i th food source position on the j th dimension for the solution space, $X_{i,j}$ is the j th dimension of the i th food source position, $X_{k,j}$ is the j th dimension of the k th food source position, φ is the scaling factor randomly produced in the range of $[-1, 1]$ and N is number of employed bees.

After all the employed bees update self-solutions, fitness values of the solutions of employed bees are calculated as follows:

$$fit_i = \begin{cases} \frac{1}{1 + f_i} & \text{if } (f_i \geq 0) \\ \frac{1}{1 + abs(f_i)} & \text{if } (f_i < 0) \end{cases} \quad (2)$$

where, fit_i is the fitness value of the solution of i th employed bee, f_i is the objective function value specific for the problem.

The onlooker bees select an employed bee in order to improve its solution by using (1). The selection probabilities of the employed bees by the onlooker bees are calculated as follows [11]:

$$p_i = \frac{0.9 \times fit_i}{fit_{best}} + 0.1 \quad (3)$$

where, p_i is being selected probability of the solution of i th employed bee, fit_i is the fitness value of the solution of the i th

employed bee and fit_{best} is the maximum fitness value of the solutions of the employed bees.

In the initialization phase of the algorithm, a food source position is produced for each employed bee using (4).

$$X_{i,j} = X_j^{min} + r \times (X_j^{max} - X_j^{min}) \quad (4)$$

$$i = 1, 2, \dots, N \text{ and } j = 1, 2, \dots, D$$

where, $X_{i,j}$ is the j th dimension of i th food source position, X_j^{min} is the lower bound for the j th dimension, X_j^{max} is the upper bound for the j th dimension, N is the number of employed bees and D is the dimensionality of the optimization problem. Also it is mentioned that if a scout bee occur (if the trial counters of the food source is higher than the limit value determined for the populations), a new food source position is produced for this scout bee by using (4).

The ABC algorithm is iterative algorithm and consists of four phases, named as initialization phase, employed bee phase, onlooker bee phase and scout bee phase, sequentially realized. The algorithm of the ABC is also given in Fig. 1.

Initialization Phase

Determine the number of employed bees or food source number (N)

Determine the *limit* value for the population

Generate food source positions for each employed bee by using (4)

Calculate the objective function values of the solutions of the employed bees

Calculate the fitness values of the solutions of the employed bees using (2)

Reset the trial counters of the food sources.

Repeat{

Employed Bee Phase

For each employed bee

Update the solution of the employed bees using (1)

Calculate the objective function value of candidate solution

Calculate the fitness value of the candidate solution

If fitness value of the candidate solution is better than the fitness value of the solution of employed bee, memorize the candidate solution and reset its trial counter, otherwise, increase its trial counter by 1

Onlooker Bee Phase

Calculate the selection probabilities using(3)

For each onlooker bee

Select an employed bee.

Update the solution of employed bee using (1)

Calculate the objective function values of the solution of the onlooker bee

Calculate the fitness value of the solution of onlooker bee.

If fitness value of the solution of the onlooker bee is better than the fitness value of the solution of employed bee, memorize the solution of onlooker bee and reset its trial counter, otherwise, increase its trial counter by 1

Scout Bee Phase

Fix the maximum content of the trial counters

If the counter with maximum content is higher than the limit, generate a new solution for this bee, calculate the objective function value of this solution, calculate the fitness value of this solution and reset trial counters of this bee.

}Until(A termination condition is met)

Fig. 1. The algorithm of ABC optimizer.

Despite the fact that only one dimension of the food source

position is updated at the each iteration in the ABC algorithm, the all dimensions of the food source position are increase convergence speed of the ABC-kNN. Because the ABC algorithm is iterative algorithm, the termination condition for the algorithm can be maximum iteration number, error rate etc.

C. K-Nearest Neighbor Algorithm

k-Nearest neighbour algorithm (kNN) is one of the most commonly used classification method. Since there is no training process in kNN, it is a simple classification method having a low computational cost compared to other classification techniques. It is especially used when there is no-knowledge of data distribution on the dataset.

In a j-dimensional dataset, k nearest sample is obtained for each sample evaluated. The distance between the samples and the sample evaluated are calculated using common distance calculation methods like Euclid, Hamming and Manhattan. Equation (5) can be given for the Euclid distance calculation method. The class of the sample evaluated can be determined by majority voting of the k nearest samples classes [12], [13].

$$d = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \tag{5}$$

In (5), p stands for the sample evaluated, q stands for any sample within the training dataset and n stands for the feature size.

D. Proposed ABC-KNN Approach

In the proposed approach, the aim is to find two centroid vectors which are equidistant to corresponding healthy or unhealthy training samples. In other words, the aim is to find two centroid vectors for both healthy and unhealthy samples which could represent the distinct class in the dataset (healthy or unhealthy). In order to achieve this aim, the dataset is divided into two subsets concerning the samples healthy and unhealthy status. Then, following steps are applied for each subset;

- 1) Random initial vectors are generated by using (4) and the objective function values given in (5) are calculated.
- 2) Repeat
- 3) The solutions of employed bees are updated using (1) in the employed bee phase.
- 4) In the onlooker bee phase, an employed bee is selected based on fitness value in order to improve its solution using (1). (This step is executed for each onlooker bee)
- 5) If scout bee occurs, a random solution is generated and assigned to the scout bee.
- 6) The best solution obtained so far is stored.
- 7) Until (A termination condition is met)

After the two centroid vectors are obtained by executing aforementioned steps, the samples are classified using kNN. Because there are more than one centroid vector for the dataset, kNN use distances of samples to the centroid vectors. If a sample is close to the centroid vector of healthy class, the sample is accepted as healthy; vice versa, if a sample is close to the centroid vector of a unhealthy class, the sample is accepted as unhealthy.

Not only the proposed approach is useful for a two class classification problem but also multi class classification

problem. The aim is to be expanded in order to obtain multiple centroid vectors.

III. RESULTS AND DISCUSSION

In the proposed approach 80 % of the EST dataset is used in order to obtain two centroid vectors corresponding to healthy or unhealthy samples from the training part of the dataset. Then, the rest part of the dataset is classified using the obtained two centroids with kNN algorithm. The proposed algorithm is repeated 50 times to provide reliability. Results of the optimal solutions for different iterations are given in Table I, and variation of the classification accuracy depending on the population size and maximum iteration number which is used as termination criterion for the ABC is given in Fig. 2.

The proposed ABC-kNN approach could be used as an alternative diagnostic accuracy tool in the CAG assessment process. The EST dataset is classified by a similar approach in which PSO and kNN is used in a related work [14]. The comparative results to related previous work are given in Table II.

It can be seen from the comparative results that ABC-kNN approach achieve more successful diagnostic accuracy then PSO-kNN approach by increasing the specificity of the classification process. And also, the sensitivity, positive predictive value and negative predictive value are almost equal for both approaches.

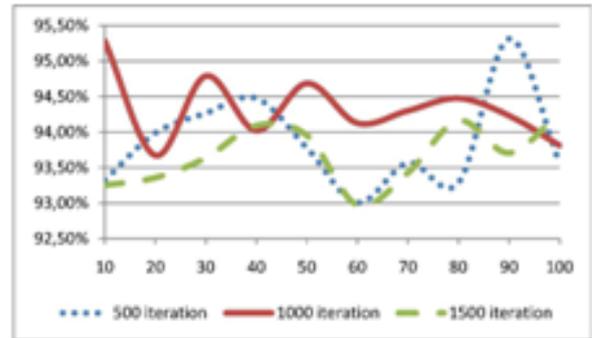


Fig. 2. Average diagnostic accuracy depending on the population size

TABLE I: OPTIMAL RESULTS OF ABC-KNN APPROACH

ITR	POP	SEN (%)	SPE (%)	PPV (%)	NPV (%)	ACC (%)
500	40	95,79	91,42	96,60	89,00	94,48
1000	10	96,83	91,68	96,65	91,75	95,28
1500	100	95,60	90,98	96,50	88,50	94,27

ITR, Maximum iteration; POP, Population size; SEN, Sensitivity; SPE, Specificity; PPV, Positive predictive value; NPV, Negative predictive value; ACC, Diagnostic accuracy.

TABLE II: COMPARATIVE RESULTS.

Method	SEN (%)	SPE (%)	PPV (%)	NPV (%)	ACC (%)
ABC-kNN	96,83	91,68	96,65	91,75	95,28
PSO-kNN	97,39	79,94	92,87	93,13	92,49

Method, The classification method; SEN, Sensitivity; SPE, Specificity; PPV, Positive predictive value; NPV, Negative predictive value; ACC, Diagnostic accuracy.

IV. CONCLUSION AND FUTURE WORKS

This paper proposes an integration of ABC algorithm and

kNN classifier for diagnosis of coronary artery disease. The experimental results show that proposed approach could be used as an alternative classification tool. Future works could include implementing ABC-kNN hybrid system to the different datasets and aim to improve its performance.

ACKNOWLEDGEMENTS

This study has been supported by Scientific Research Project of Selçuk University.

REFERENCES

- [1] R. Gibbons, G. Balady, J. T. Bricker *et al.*, "ACC/AHA guideline update for exercise testing: summary article. A report of the American college of cardiology/American heart association task force on practice guidelines (Committee to update the 1997 exercise testing guidelines)," *J Am Coll Cardiol*, vol. 40, no. 8, pp. 1531–1540, 2002.
- [2] J. A. San Roman, I. Vilacosta, J. A. Castillo, M. J. Rollan, M. Hernandez, V. Peral *et al.*, "Selection of the optimal stress test for the diagnosis of coronary artery disease," *Heart*, vol. 80, no. 4, pp. 370–376, 1998.
- [3] T. Thom, N. Haase, W. Rosamond, V. J. Howard, J. Rumsfeld, T. Manolio *et al.*, "Heart disease and stroke statistics–2006 update: A report from the American heart association statistics committee and Stroke statistics subcommittee," *Circulation*, vol. 113, pp. e85–e151, 2006.
- [4] I. Maglogiannisa, E. Loukisb, E. Zafiropoulosb, and A. Stasisb, "Support vectors machine-based identification of heart valve diseases using heart sounds," *Computer Methods and Programs in Biomedicine*, vol. 95, no. 1, pp. 47–61, 2009.
- [5] S. Choi, "Detection of valvular heart disorders using wavelet packet decomposition and support vector machine," *Expert Systems With Applications*, vol. 35, no. 4, pp. 1679–1687, 2008.
- [6] Z. Zhidong, "Noninvasive diagnosis of coronary artery disease based on instantaneous frequency of diastolic murmurs and SVM," in *Proc. IEEE Engineering in Medicine and Biology 27th Annual Conference, Shanghai, China*, pp. 5651–5654, 2005.
- [7] M. Kukar, C. Groselj, I. Kononenko, and J. J. Fettich, "An application of machine learning in the diagnosis of ischaemic heart disease," in *Proc. 10th IEEE Symposium on Computer-Based Medical Systems (CBMS'97)*, pp. 70, 1997.
- [8] P. Lapuerta, S. P. Azen, and L. Labree, "Use of neural networks in predicting the risk of coronary artery disease," *Computers and Biomedical Research*, vol. 28, no. 1, pp. 38–52, 1995.
- [9] D. Karaboga, B. Gorkemli, C. Ozturk, and N. Karaboga, "A comprehensive survey: artificial bee colony (ABC) algorithm and applications," *Artificial Intelligence, Review*, doi:10.1007/s10462-012-9328-0, 2012.
- [10] D. Karaboga, "An idea based on honey bee swarm for numerical optimization," Technical report, Computer Engineering Department, Engineering Faculty, Erciyes University, 2005.
- [11] D. Karaboga, C. Ozturk, N. Karaboga, and B. Gorkemli, "Artificial bee colony programming for symbolic regression," *Information Sciences*, vol. 209, pp. 1–15, 2012.
- [12] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern classification," 2nd ed. *John Wiley and Sons*, New York, 2001.
- [13] G. Shakhnarovich, T. Darrell, and P. Indyk, *Nearest-Neighbor Methods in Learning and Vision*. MIT Press 2005.
- [14] İ. Babaoglu, O. Fındık, E. Ülker, and N. Aygül, "A novel hybrid classification method with particle swarm optimization and k-nearest neighbor algorithm for diagnosis of coronary artery disease using exercise stress test data," *International Journal of Innovative Computing, Information and Control*, vol. 8, no. 5(b), pp. 3467–3475, 2012.