Journal of Industrial and Systems Engineering
Vol. 14, Special issue: 18th International Industrial Engineering Conference
Winter (January) 2022, pp. 40-47

JISE

17 November 2021
Iran Institute of Industrial Engineering

Predicting coronary artery diseases using effective features selected by Harris Hawks optimization algorithm and support vector machine

Sarina Maleki[1], Yahia Zare Mehrjerdi[1*], Davoud Shishebori[1], Masoud Mirzaei[2]

[1]*Department of Industrial Engineering, Technical Engineering Faculty, Yazd University, Yazd, Iran*
[2]*Disease Modeling Center of Shahid Sadoughi University of Medical Sciences, Yazd, Iran*

sarinamaleki1398@gmail.com, mehrjerdyazd@gmail.com, shishebori@yazd.ac.ir, masoud_mirzaei@hotmail.com

Abstract

With 17 million annual deaths, cardiovascular diseases are the leading cause of mortality across the world with coronary artery disease (CAD) as the most prevalent one. CAD is the leading cause of death in industrial countries and at the same time is rapidly spreading in the developing world. Thus, the development and introduction of machine learning methods for the accurate diagnosis of heart diseases, especially CAD, have been an important debate in recent years in order to overcome relevant problems. The aim of this paper was to propose a model for enhancing CAD prediction accuracy. It sought a framework for predicting and diagnosing CAD using the features selection of Harris Hawks Optimization algorithm (HHO) and Support Vector Machine (SVM). The heart disease data set of Cleveland hospital available in the University of California Irvine (UCI) was used as the studied data set. It included 303 cases. Each case had 14 features with the final medical status of cases (CAD or normal case) as one of the features where 165 and 138 cases were diagnosed as CAD and normal, respectively. The results of this study revealed that HHO could enhance CAD diagnosis accuracy.

Keywords: Coronary artery diseases, feature selection, Harris Hawk optimization algorithm, support vector machine

1-Introduction

Rapid health transition, demographic changes, aging, and rapid changes to life styles, besides other socio-economic evolutions have led to the ever-increasing involvement in non-communicable diseases and induced disabilities and mortality (Reddy, 2002). Cardiovascular diseases include those vascular system diseases affecting cerebral, heart and peripheral circulations (Ndindjock et al., 2011). The problem is that heart diseases have been the leading cause of mortality in the past 10 years.

*Corresponding author
ISSN: 1735-8272,

As the most prevalent heart disease CAD is the leading cause of mortality in industrial countries and is rapidly spreading in the developing world (Ndindjock et al., 2011). Several methods, including cardiac stress test, echocardiogram, nuclear heart scan, and angiography are used to diagnose CAD (Khosravanian and Ayat, 2015) and (Nahar et al., 2013).

These are generally error-prone diagnostic tests and are very costly, time-consuming and troublesome for patients (Negahbani et al., 2015). Thus, the development and introduction of machine learning methods for the accurate diagnosis of heart diseases, especially CAD, have been an important debate in recent years in order to overcome relevant problems (Giri et al., 2013) and (Engelbrecht, 2007). CAD is a serious disease with a significant mortality rate. It accounts for 31% of global annual deaths (17.9 million deaths per year) (Nasarian et al., 2020)[8]. Therefore, this paper aims to enhance CAD diagnosis accuracy using supervised learning machine methods, including SVM. This goal will be achieved through Harris Hawks optimization algorithm (HHO) which is inspired by Harris hawks hunting style. The aim is to derive a subset out of total features of CAD data set using HHO meta-heuristic algorithm. The selected features diagnose CAD more accurately compared to the case where all features are used for diagnosis. The increased number of parameters makes diagnosis process very difficult even for an expert medical specialist. Therefore, this paper tries to solve this problem and assist CAD diagnosis using more effective features selected by HHO feature selection.

2-Background

There are many studies on CAD diagnosis. Rani analyzed CAD data using neural networks. He employed a single-layer neural network and a multi-layer one on the studied data set and obtained the accuracy values of 87% and 83%, respectively (Rani, 2011). Dezhaloud et al used Binary Grasshopper Optimization algorithm and k-Nearest Neighbor Machine Learning and diagnosed CAD in the data set of 270 patients with an accuracy of 89.8% (DezhAloud, 2020). Vula et al used Bayesian Networks to diagnose heart diseases. Their proposed method identified normal cases with an accuracy of 91% (Vila-Francés et al., 2013). Moloud Abdar et al diagnosed CAD with an accuracy of 93.08% using Genetic Algorithm (GA), Particle Swarm Optimization Algorithm, and Support Vector Machine (SVM) (Abdar et al., 2019). It can be argued from this background that meta-heuristic algorithms have not been in the area of focus for reducing the number of features and selecting more effective ones while this initiative can enhance diagnosis accuracy. Furthermore, no study has been conducted on the combined application of HHO and SVM for feature selection. Therefore, this study aims to enhance CAD diagnosis accuracy through HHO feature selection. The data set of Cleveland hospital CAD patients was used as the studied data set. It was provided from the machine learning data sets of the University of California (UCI). It includes 303 cases. Each case has 14 features and 165 and 138 cases were diagnosed as CAD and normal, respectively. Table 1 shows the features.

Table 1. Cleveland data set features

No.	Name	Description	Range
1	Age	Age	29-77
2	Sex	Sex	male-female
3	Cp	Chest Pain Type	0-3
4	trestbps	Resting Blood Pressure (mmHg)	94-200
5	chol	Serum Cholesterol	126-564
6	fbs	Fasting Blood Sugar	0-1
7	restecg	Resting electrocardiogram	0-2
8	thalach	Max. heart beat	71-202
9	exang	Exercise-induced angina	0-1
10	oldpeak	Exercise-induced ST depression compared to resting state	0-6.2
11	slope	slope of ST wave peak	0-2
12	ca	Number of large veins detected by fluoroscopy	0-4
13	thal	Thallium Scientigraphy	0-1

3-Method of study

3-1-Steps of the proposed method

The general method of this study is about enhancing CAD diagnosis accuracy using HHO feature selection and machine learning classification algorithms (SVM). Data-pre-processing is conducted in the first step. It includes data selection, data cleansing, data transformation, and data normalization stages. Following data pre-processing, HHO feature selection is used to reduce dimension and to select effective features. In this way, a sub set of the effective features of CAD diagnosis is selected. Next, the newly built data set, with limited features and reduced dimension compared to total features, is introduced to SVM. This study uses CAD patients' data, available in the machine learning data sets of the University of California, and adopts machine learning algorithms, including SVM, to predict and diagnose CAD. In addition, it selects more effective features by feature selection methods such as HHO to enhance the precision level of models. Finally, the results of evaluation parameters i.e. "accuracy" and "SVM performance" are obtained for "with feature selection" and "without feature selection" methods and compared with each other.

3-2- Evaluation and validity indicators

This study used the following valid measures to evaluate and compare the efficiency of different machine learning models in predicting and diagnosing CAD (Glaros and Kline, 1988).

3-2-1- Confusion Matrix

As far as the classification of a data set using machine learning classification methods concerns, the aim is to classify and identify classes with the highest possible accuracy. In some problems, it is strongly important to accurately identify the cases of a given class. Consider a study where the aim is to identify individuals with a particular dangerous disease. Assume that the patients are susceptible to death and they need a special drug. In this condition, it is of a high importance to accurately differentiate the patients. This means that any mistake in differentiating normal cases is ignorable while the same is not true for differentiating a patient as a normal case. In other words, it is expected to detect all patients with no missed case, even in the expense of classifying a normal case as a patient. In the conditions where the accuracy of detecting a given class governs the overall accuracy, the concept of confusion matrix assists us. Consider above example again. Assume that the inclusion of a case in the patient class is considered to be positive, and the contrast situation is considered to be negative. In reality, each case belongs either to the positive class or to the negative class. On the other hand, any and all classification algorithms will classify each case within one of these classes. Therefore, the probable states for each case will be as the following:

- True Positive (TP): the case belongs to the positive class and is detected as a member of this class.
- False Negative (FN): the case belongs to the positive class and is detected as a member of the negative class.
- True Negative (TN): the case belongs to the negative class and is detected as a member of this class.
- False Positive (FP): the case belongs to the negative class and is detected as a member of the positive class.

After running the classification algorithm, classification performance may be evaluated in accordance with table 2, considering above definitions:

Table 2. Concept of confusion matrix

Identified Label	
Negative	Positive
FP	TP
TN	FN

3-2-2- Accuracy

Accuracy equals to the ratio of truely classified data to total data which is stated in percent:

Accuracy=TP+TN/TP+FN+FP+TN (1)

3-2-3- Sensitivity

Sensitivity refers to true positive rate (TPR) and is calculated from the following relation:

Sensitivity=TP/TP+FN (2)

3-2-4- Specificity

Specifity refers to True Negative Rate (TNR) and is calculated from the following relation:

Specificity(TNR)=TN/TN+FP (3)

3-2-5- Precision

Precision shows the proportion of positive samples which are actually positive samples

Precision=TP/TP+FP (4)

3-2-6- F-Score

F-score is a measure used to evaluate the performance of classification algorithms. It is composed of Recall and Precision parameters and is the harmonic mean of them:

$$F-Score = 2 \times \frac{(\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})} \quad (5)$$

3-2-7- Mean Square Error (MSE)

MSE is a fitness function or an objective function. It is an indicator of absolute error between the simulated and observed variable. It ranges from zero to infinity. The higher the value, the better is the simulation. The optimum value would be zero. It is stated by the following equation:

$$MSE = \frac{1}{n} \times \sum_{i=1}^{n} [(x_{imeas} - x_{ipred})^2] \quad (6)$$

n, x_{ipred} and x_{imeas} are number of measured variable, value of predicted variable, and value of measured variable, respectively.

3-3- Objective function of the feature selection of HHO and Bat algorithms

$X=(x_1, x, \ldots x_d)$

F(x)= Accuracy(x)

Accuracy=TP+TN/ TP+FN+FP+TN

The aim of this study is to enhance CAD diagnosis accuracy. According to above relation, the inputs of the objective function are HHO-selected features and the output of it is the accuracy. In addition, x indicates HHO-selected features. For any iteration, HHO introduces different features to the objective function and calculates the accuracy. Finally, when iterations completed, those HHO-selected features which maximize accuracy are introduced as the effective features of CAD diagnosis.

3-4- HHO feature selection

HHO is a population-based nature-inspired algorithm. The main idea behind HHO is the cooperative behavior and chasing style of Harris hawks in nature called surprise pounce (Heidari et al., 2019). In this intelligent strategy, several hawks cooperatively pounce a prey from different directions in an attempt to surprise it. Harris hawks can reveal a variety of chasing patterns based on the dynamic nature of scenarios and escaping patterns of the prey. In 1997, Louise Lefebvre introduced an approach to measuring the intelligence quotient of birds based on innovations observed in their feeding behaviors. According to his studies, hawks can be classified among the most intelligent birds in nature. Harris hawks are well-known prey-hunting birds. They could be found in relatively stable groups in the south half of Arizona, America. To select features using HHO, initial values were allocated to HHP parameters as per table 2:

Table 2. HHO parameters for feature selection

HHO algorithm parameters	**Hawks population**	**number of iteration**	**Dimensions (D)**	β
value	50	100	54	1.5

3-5- CAD Diagnosis using HHP and SVM (Cleveland data set)

CAD was predicted by SVM through two methods: with and without HHO feature selection. The results and evaluation scores of both methods are shown in the following graph. This graph shows the values of the objective functions for 100 iterations of HHO performed to select effective parameters of CAD diagnosis (figure 1).

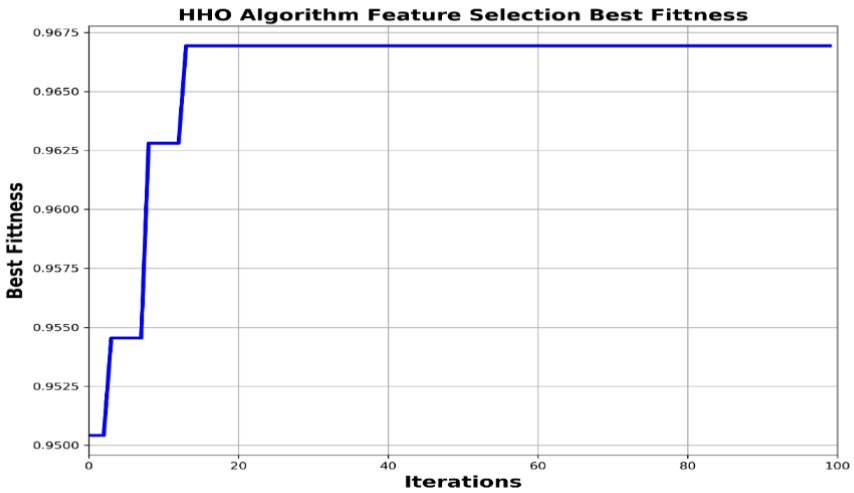


Fig. 1. HHO convergence for 100 iterations performed to select effective features (SVM) (Cleveland data set)

Table 3. SVM predictions of CAD using the features selected by HHP and total features (Cleveland data set)

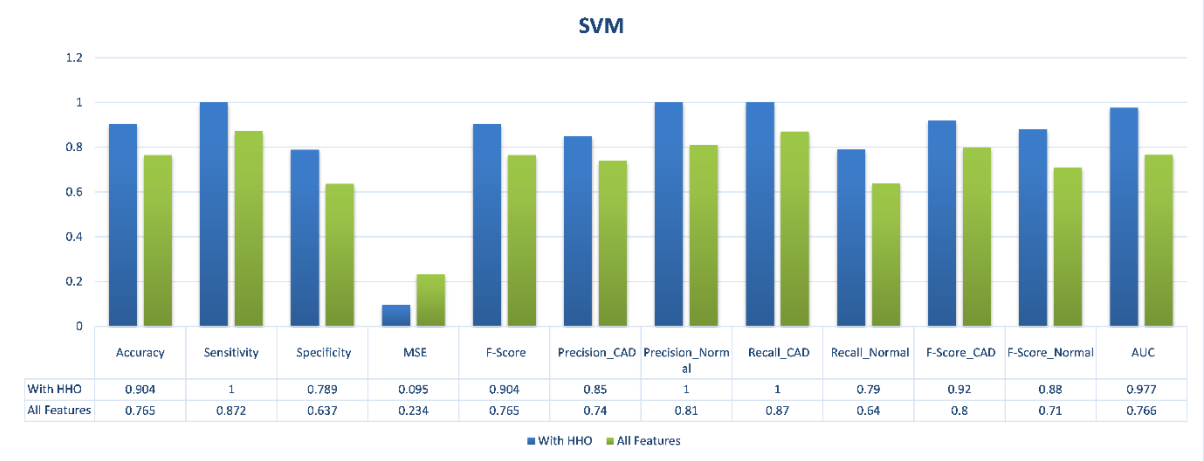
Parameter	HHO optimized values	predicted by all features
Accuracy	0.904	0.765
Sensitivity	1.0	0.872
Specificity	0.789	0.637
MSE	0.095	0.234
F-Score	0.904	0.765
Precision (patient class)	0.85	0.74
Precision (normal class)	1.00	0.81
Recall (patient class)	1.00	0.87
Recall (normal class)	0.79	0.64
F-Score (patient class)	0.92	0.80
F-Score (normal class)	0.88	0.71
ROC (AUC)	0.977	0.766
number of features selected by HHO	6	13
features select by HHO	1 2 4 8 10 11	-

4-Conclusion

The aim of this study was to enhance CAD diagnosis accuracy using four supervised machine learning techniques _ namely: Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Tree (DT), and K-nearest neighbor (KNN). It selected a subset out of total features available in CAD data set using Bat meta-heuristic Optimization algorithm. The selected features diagnosed CAD more effectively compared to all features. The data set used in this study was for CAD patients of Cleveland hospital (UCI). It included 303 cases. Each case had 14 features with the final medical status of cases (CAD patient or normal case) as one of the features where 165 cases were CAD and 138 were normal cases.

4-1- Comparison of SVM results (Cleveland data set)

According to the following figure, by selecting 6 features out of 13 features, HHO could rise CAD diagnosis accuracy where accuracy, sensitivity, specificity, F-Score, Precision of patient class, Precision of normal class, Recall of patient class, Recall of normal class, F-Score of patient class, F-Score of normal class, and AUC raised by 14%, 13%, 15%, 14%, 11%, 19%, 13%, 15%, 12%, 17%, and 21%, respectively. In addition MSE reduced by 14%.



	Accuracy	Sensitivity	Specificity	MSE	F-Score	Precision_CAD	Precision_Normal	Recall_CAD	Recall_Normal	F-Score_CAD	F-Score_Normal	AUC
With HHO	0.904	1	0.789	0.095	0.904	0.85	1	1	0.79	0.92	0.88	0.977
All Features	0.765	0.872	0.637	0.234	0.765	0.74	0.81	0.87	0.64	0.8	0.71	0.766

Fig. 2. Comparison between SVM results with and without HHO-selected features (Cleveland data set)

References

Abdar, M., Książek, W., Acharya, U. R., Tan, R. S., Makarenkov, V., & Pławiak, P. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. *Computer methods and programs in biomedicine*, *179*, 104992.

DezhAloud, N. (2020). Diagnosis of Heart Disease Using Binary Grasshopper Optimization Algorithm and K-Nearest Neighbors. *Journal of Health Administration*, *23*(3), 42-54.

Engelbrecht, A. P. (2007). *Computational intelligence: an introduction*. John Wiley & Sons.

Glaros, A. G., & Kline, R. B. (1988). Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model. *Journal of clinical psychology*, *44*(6), 1013-1023.

Giri, D., Acharya, U. R., Martis, R. J., Sree, S. V., Lim, T. C., VI, T. A., & Suri, J. S. (2013). Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform. *Knowledge-Based Systems*, *37*, 274-282.

Heidari, A. A., Mirjalili, S., Faris, H., Aljarah, I., Mafarja, M., & Chen, H. (2019). Harris hawks optimization: Algorithm and applications. *Future generation computer systems*, *97*, 849-872.

Khosravanian, A., & Ayat, S. S. (2015). Presenting an intelligent system for diagnosis of coronary heart disease by using Probabilistic Neural Network.

Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P. (2013). Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems with Applications*, *40*(4), 1086-1093.

Nasarian, E., Abdar, M., Fahami, M. A., Alizadehsani, R., Hussain, S., Basiri, M. E., ... & Sarrafzadegan, N. (2020). Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach. *Pattern Recognition Letters*, *133*, 33-40.

Ndindjock, R., Gedeon, J., Mendis, S., Paccaud, F., & Bovet, P. (2011). Potential impact of single-risk-factor versus total risk management for the prevention of cardiovascular events in Seychelles. *Bulletin of the World Health Organization, 89*, 286-295.

Negahbani, M., Joulazadeh, S., Marateb, H. R., & Mansourian, M. (2015). Coronary artery disease diagnosis using supervised fuzzy c-means with differential search algorithm-based generalized Minkowski metrics. *Peertechz Journal of Biomedical Engineering, 1*(1), 006-014.

Rani, K. U. (2011). Analysis of heart diseases dataset using neural network approach. *arXiv preprint arXiv:1110.2626*.

Reddy, K. S. (2002). Cardiovascular diseases in the developing countries: dimensions, determinants, dynamics and directions for public health action. *Public health nutrition, 5*(1a), 231-237.

Vila-Francés, J., Sanchis, J., Soria-Olivas, E., Serrano, A. J., Martinez-Sober, M., Bonanad, C., & Ventura, S. (2013). Expert system for predicting unstable angina based on Bayesian networks. *Expert systems with applications, 40*(12), 5004-5010.