Al-Rafidain Journal of Computer Sciences and Mathematics (RJCM)

www.csmj.mosuljournals.com

Hybridization of Swarm for Features Selection to Modeling Heart Attack Data

Omar Shakir Hasan [1],* Ibrahim Ahmed Saleh [2]

1computer science, College of Computer Science and Mathematics, Mosul University, Mosul , Iraq
2software engineering, College of Computer Science and Mathematics, Mosul University, Mosul , Iraq
Email: omarshakir06@gmail.com

Article information

Article history:
Received : 24/4/2022
Accepted : 21/6/2022
Available online :

Abstract

Predicting heart attacks using machine learning is an important topic. Medical data sets contain different features, some of which are related to the target group for prediction and some are not. In addition, the data sets are excessively unbalanced, which leads to the bias of machine learning models when modeling heart attacks. To model the unbalanced heart attack data set, this paper proposes the hybridization of Particle swarm optimization (PSO), BAT, and Cuckoo Search (CS) to select the features and adopt the precision for minority classes as a fitness function for each swarm to select the influential features. In order to model the data, set in which the features were selected, it was proposed to use the boosting (Catboost) as a classifier for predicting heart attacks. The proposed method to select features has been compared with each of the three swarms, and the Catboost algorithm has been compared to traditional classification algorithms (naive Bayes, decision trees). The study found that the proposed method of hybridization of the results of the (PSO, BAT, and BCS) algorithms in selecting features is a promising solution in the field of selecting features and increases the accuracy of the system, and that traditional machine learning models are biased in the case of unbalanced data sets and that selecting the important features according to the target class has an impact on the performance of the models, In addition, the definition of hyperparameters reduces the bias of the selected model. The final model achieved an overall accuracy of 96% on the Accuracy scale and 56% on the Precision scale for the minority class

Keywords:
Machine Learning, Imbalance Data, Cuckoo Search, Particle swarm optimization, BAT algorithm, Naive Bayes, Decision Trees, Catboost

Correspondence:
Author : Omar Shakir Hasan
Email: omarshakir06@gmail.com

I. INTRODUCTION

In modern society, heart disease is one of the world's fatal diseases and will become the world's largest disease burden [1]. Heart disease includes coronary heart disease (heart attack), congestive heart failure, stroke, peripheral artery disease, carotid artery disease, and aortic disease. [2]. With the rapid development of computer technology and artificial intelligence (AI), machine learning technology has opened up

new ideas for risk assessment of disease prediction. Because AI systems need to have the ability to acquire knowledge by themselves, that is, the ability to extract patterns from raw data, this ability is called machine learning (ML)[3][4].

The introduction of machine learning allows computers to solve many problems related to the real world and make seemingly subjective decisions. Accumulated a large amount of data from heart patients in the electronic health records (EHR). Still, the busy clinical environment makes the integration and effective use of these data extremely challenging, so the data itself does not better serve clinical decision-making ability [5][6]. In addition, many studies based on biomedical data come from conventional assumptions, that

is, to explore the impact of selected preselected variables on cardiovascular phenotypes [7]. In contrast, AI-based methods can be used under hypothetical conditions. Multiple variables drive data mining and problem discovery to select the similarities and differences of phenotypes between patients[a8]. Therefore, standardizing clinical diagnosis, improving existing treatment methods, finding new drug targets, and achieving data-driven high-quality care at a higher rate are important implementation measures to promote innovation in the medical field [9].

Research studies attempt to model medical record data based on the above starting point. To model those records, data are analyzed and processed. Modeling algorithms and statistical methods are used to reduce the error between the expected results and the real results. The potential of the accumulated data and risk variables are determined. The complex and non-linear effects between these variables are explored, and a heart attack prediction model is created based on data sets [10][11]. The assessment of heart attack risk relies on various risk factors for cardiovascular disease to predict an individual's likelihood of having an acute heart attack [12]. In this way, the corresponding intervention measures are taken to reduce the influence of risk factors, prevent and reduce the occurrence of such clinical events in a timely manner, and improve the health of the whole community. To find the appropriate model for predicting heart attacks [13]. In this paper, medical data sets and previous literature surveys were analyzed, hybridization of the results of the (PSO, BAT, and CS) algorithms and adopt the precision for minority classes as a fitness function for each swarm to select the important features. In order to model the data set in which features were selected, proposed Catboost as a classifier for heart attack prediction. The proposed method to select features has been compared with each of the three swarms, and the Catboost algorithm was compared with traditional classification algorithms (naive Bayes, decision trees)

The paper has been organized as follows, first surveying the previous literature, then presenting the theoretical framework of the algorithms specified and the proposed methodology, which included displaying the data set and selecting the features using swarms algorithms in addition to the proposed method (HSFS) then applying the selected machine learning algorithms. In the end, testing and comparing the models based on the methods of selecting features using the accuracy scale and the precision.

2. Related works

Kim et al [14], used (KNHANES-VI) dataset and neural network feature correlation analysis (NN_FCA) method has been proposed. Since (FCA) includes two stages, the first stage is feature selection, and the second stage is feature correlation. Then used is a neural network algorithm for classification. This method improved the neural network algorithm performance for predicting heart disease. Accuracy reached 85.70 %.

Kasbe et al [15], the authors suggested a fuzzy expert system for predicting heart disease. It consisted of three main steps fuzzification, rule base, and defuzzification. for defuzzification, the centroid technique had used. The system contains 13 input parameters and one output parameter, using the heart disease dataset UCI repository. The system is very easy in usability, and the patient can use it by themselves. The accuracy of this method achieved a 93.33%.

Malav et al [16], in this study, K means, and artificial neural networks have been hybridized to build a heart disease prediction model, applied to the heart disease UCI dataset. They reached an accuracy of 97%. The results showed that the hybrid systems were superior to the traditional machine learning algorithms.

Kamboj et al [17], the authors have compared the performance of machine learning algorithms to predict heart disease such as (SVM, KNN, Naive Bayes, Random Forest Classifier, Logistic Regression) on a heart disease UCI dataset. The study concluded that KNN is the best classifier with an accuracy of 87% compared to the rest of the specified algorithms.

Riazet et al [18], the authors suggested building a predictive system for predicting heart disease at an early stage using artificial neural networks. They used PCA for feature extraction. PCA improved the results to an accuracy rate of 97.7% compared to 94.7.

Shah et al [19], the Cleveland dataset from the UCI repository has been used, which comprised 303 states and 76 features. Apply pre-processing on this dataset, such as processing missing values and removing the noise. Used only 14 most important features, then machine learning supervised algorithms on this dataset applied such as KNN, Decision trees, random forest, naïve Bayes. The KNN algorithm achieved the highest accuracy, equal to 90.78.

Siva et al [20], this paper used the heart disease dataset from the UCI repository, then made data pre-processing features selection and applied these features on the hybrid random forest with a linear model for predation heart disease. The accuracy of this method achieved 92%.

Walaa Adel Mahmoud et al [21], the authors have used the Framingham dataset, this dataset is unbalanced. The imputation means method has been used to handle missing data and outlier data values. The authors proposed using

different classifier algorithms such as (k nearest neighbours, support vectors machine, decision tree, linear regression random forest). The accuracy reached 83.95,84,5,84.89, and 85.05% for the as (k nearest neighbours, support vector machine, decision tree, linear regression random forest) algorithms respectively.

Table 1- Previous research to modeling heart attack

No	Paper	Year	Method	Accuracy
1	Paper [14]	2017	(NN_FCA)	85.70 %.
2	Paper [15]	2017	Fuzzy Expert System	93.33%.
3	Paper [16]	2017	K means, and Artificial Neural Networks	97%.
4	Paper [17]	2020	SVM, KNN, Naive Bayes, Random Forest, and Logistic Regression	KNN with higher accuracy of 87%
5	Paper [18]	2020	Artificial Neural Networks. and PCA	97.7%
6	Paper[19]	2020	KNN	90.78
7	Paper[20]	2020	Hybrid Random Forest with a Linear Model	92%.
8	Paper[21]	2021	KNN, SVM ,DT , L R and Random Forest)	Random forest is higher accuracy rechead of 85.05%

3. Theoretical Framework

3.1. Feature selection based on binary swarm

Feature selection is to find a targeted subset from the feature set of the original data to carry the most effective classification information. Feature selection aims to select as few feature subsets as possible according to a certain algorithm to achieve the best possible classification [22]. Three binary swarms were used in this study to select features .

3.1.1. Binary particle swarm optimization (PSO)

The main steps of the applied feature selection algorithm are shown in Figure (3).The process begins with generating an array of particles with random locations in the search space In the next step, the pbest and gbest are calculated at the end of each iteration. If the solution found by the particle is higher than the previous pbest, this solution will be the new pbest for that particle and the best pbest among all the particles [23].

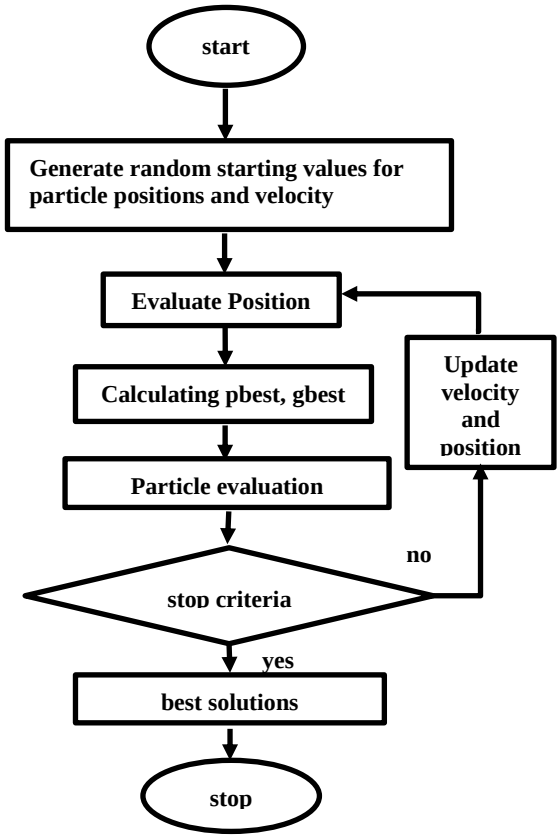

Figure 1- Particle swarm optimization flowchart

3.1.2. Binary bat

The algorithm uses basic rules that bats use echolocation in the sensing space. Bats can differentiate between danger and food [26].

Bats (binary) fly randomly, quickly, in a position, and at a fixed frequency, with different wavelengths and loudness, to search for prey. Bats (binaries) can automatically adjust the wavelength (or frequency) of their emitted pulses and adjust the pulse emission rate depending on the proximity of their target. Loudness can vary in many ways [24]. Figure (2) shows the stages of finding features using the bat algorithm [25]

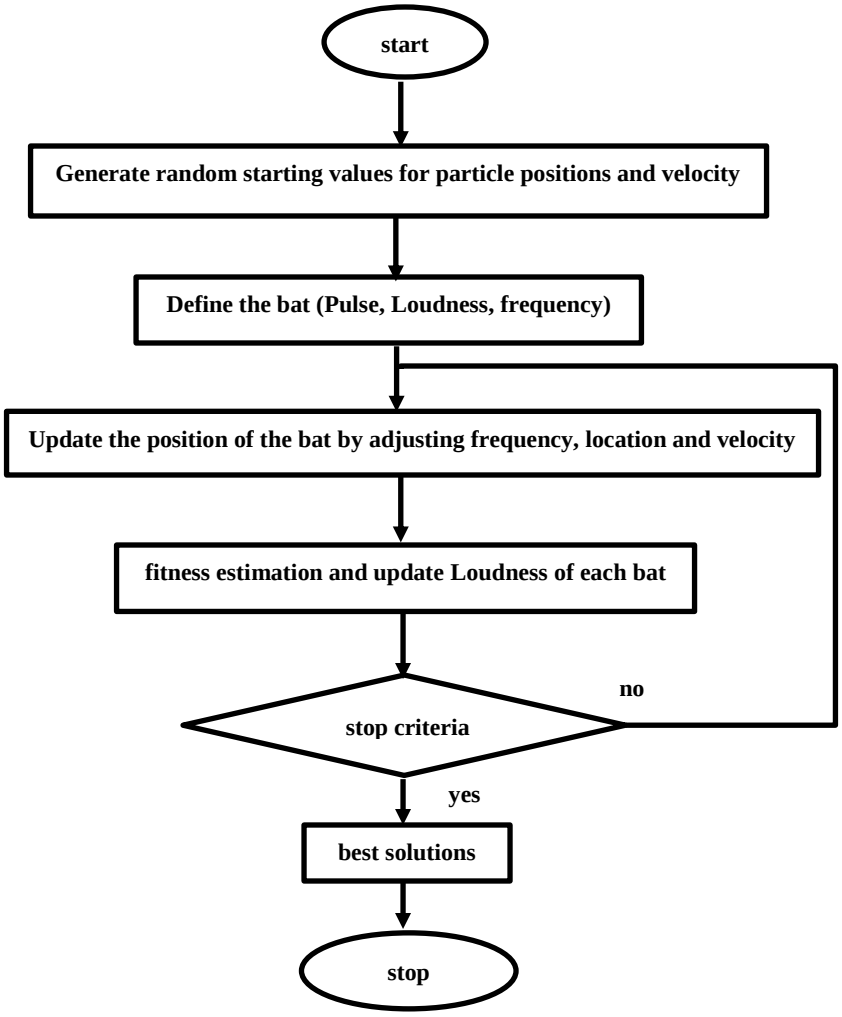

Figure 2- Binary BAT flowchart

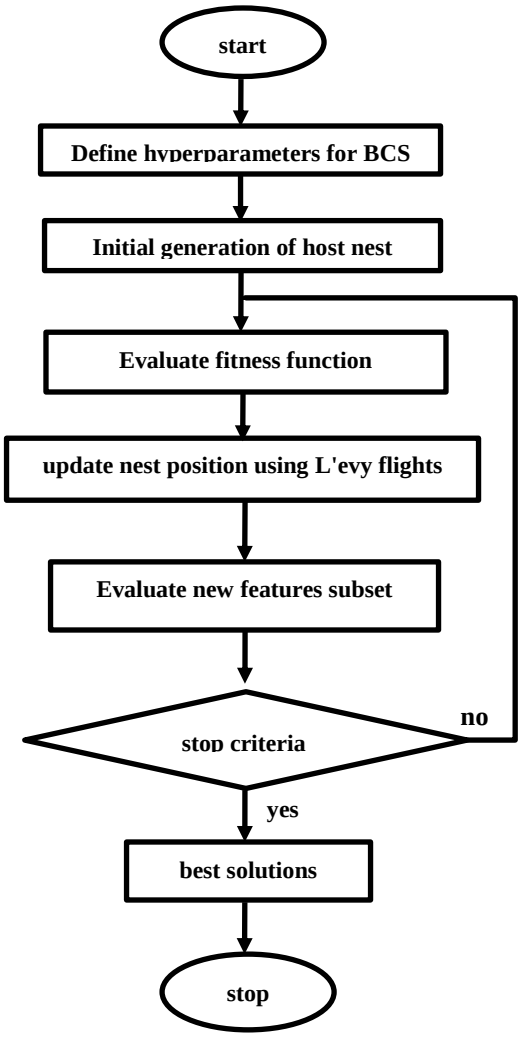

Figure 3- Binary Cuckoo Search flowchart

3.1.3. Binary cuckoo search (BCS)

Another swarm used to select the features is BCS to find the best features. Each host nest is specified algorithmically as an agent carrying a single egg (unique dimension problem) or several eggs (multidimensional problem). CS begins by randomly arranging the nest population in the search space. In each algorithm iteration, the nests are updated using random walk via L'evy flights [26]. Figure (3) shows the stages of finding features using the bat algorithm.

3.2. Modeling Heart Attack

3.2.1. Naive bayes

The models built using the Naive Bayes algorithm are considered the simplest models. It does not contain any parameters because the probability of the dependent variable is calculated from the probability of the event. To build a naive Bayes model for classifying bank loans, (GaussianNB) has been used [27]. The following is equation of Naive Bayes

$$p(A/B) = \frac{P(B/A).\ P(A)}{P(B)} \quad (1)$$

p(A | B): the probability of event A for a given event B
p(B | A): Given event A, the probability of event B
p(A): the probability of event A
p(B): the probability of event B

3.2.2. Decision trees

A decision tree is one of the supervised machine learning methods used in "Classification" and "Regression" problems. The principle of dividing into nodes from top to bottom, the data set is divided into smaller and smaller subgroups until reaching the target nodes (class), where it starts from the root node that contains all records, and then divides according to the "Class Label" column (which is the column that classification is based on it), as for its algorithms, it has many types according to the data set. In this study, type DT C4.5 was used [28] [29]. The feature selection for each node depends on the following measurements:

Information entropy: Information entropy is an index to measure the degree of disorder of elements.

$$E(S) = \sum_{I=1}^{C} -P_I \, Log_2 \; P_I \tag{2}$$

Where S is the total number of features, c is the number of classes and p(i) indicates the probability of records belonging to that class.

Information gain: Information gain measures the change in information entropy between independent attributes.

$$Gain(S, X) = Entropy(S) - Entropy(S, X) \tag{3}$$

Where Gain((S, X)) is the information gain of feature X. Entropy(S) is the information entropy of the entire dataset, and entropy(S, X) is the information entropy of feature X.

Classification error:

$$\text{Classification error}(T) = 1 - \left[p\left(\frac{i}{t}\right)\right] \tag{4}$$

p(i/t) indicates the probability of records belonging to that class

3.2.3. CatBoost

CatBoost is a new gradient-boosting algorithm introduced by Prokhorenkova et al. (2018), and its performance has been proven to be quite exciting compared to another boosting algorithm. In particular[30]. CatBoost splits a given data set into random permutations. By default, CatBoost creates four random permutations. Randomness can stop modifying our model [31]. The following mathematical formula can represent it:

$$X_k^i = \frac{\sum_{j=1}^{j}\left[X_j^i = X_k^j\right] y_j + \alpha P}{\sum_{j=1}^{n}\left[X_j^i = X_k^j\right] + \alpha} \tag{5}$$

Where α is the corresponding weight, P denotes the prior value $x_k = (X_k^1 , , , , , , , , , X_k^m$ is the random vector of m features and y_k ε R denotes the corresponding label

4. Suggested methodology

To find the appropriate model for predicting heart attacks, a data set containing patient records were used as input and whether or not a heart attack occurred as a final output. The first stage is the data initialization process, which includes cleaning the data in the event of anomalies, missing values, or categorical data, and dividing the data set into two groups, a training group and a test group with a ratio of (80:20) respectively. The stage of preparing the data for training includes selecting the features. The concept of swarms or the so-called (binary swarms) was used, and each method was applied separately, and the effect of the outputs of each method was measured.

Three swarms were applied (PSO, BAT, BCS), and compared the results among them were. a new hybrid swarm method was proposed called (HSFS)and compared the proposed method with each swarm separately. Feature selection is applied to the training data. After the swarms are applied, features are determined in the training and test data for each method on a separate data set in addition to the original data set. After creating the total and sub-datasets, the selected algorithms (naive Bayes, decision trees, and CatBoost) are trained to build models. Then measure the performance of each model concerning the method of selecting data and measure the system's performance as a whole, in addition to finding the precision of each classification class and comparing them. After the models are built and tested using the performance measures relevant to the study objective (Accuracy, Precision), the comparison is made, and the proposed methods' importance and applicability are determined. Figure (4) represents the framework proposed in this paper.

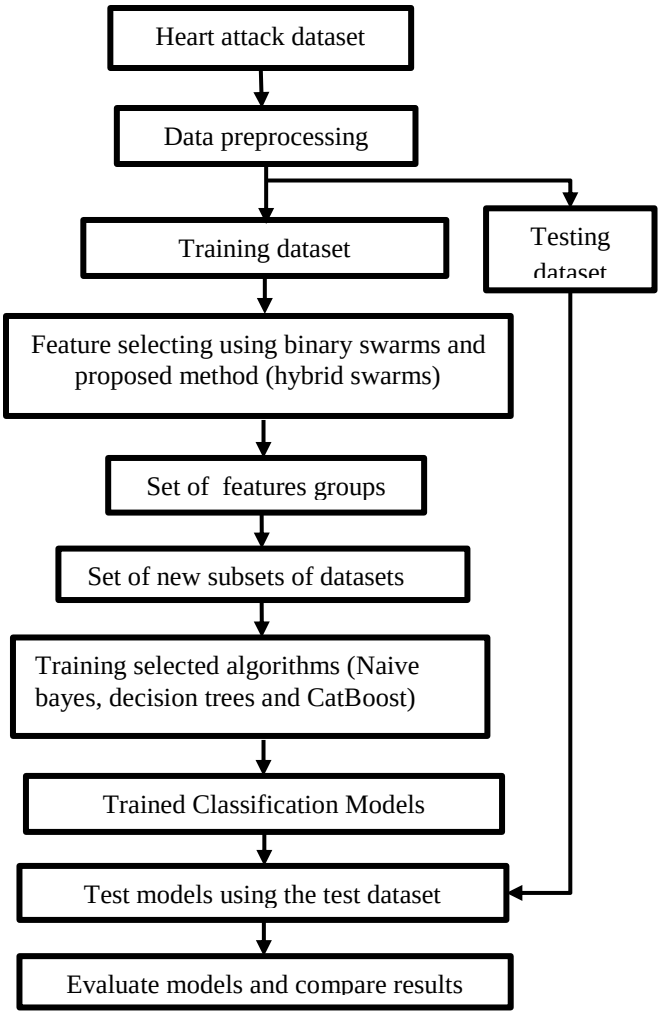

Figure 4- The framework for the proposed system

4.1. Heart Attack Dataset

In this paper, NHANES data were used for information on adults over 18 years. The dataset contains 37,080 records of diverse individuals. 1,300 people have coronary heart disease (heart attack) and 35780 non-coronary heart diseases. Each person has 50 features and one target value (Non-CHD and CHD). The data set is unbalanced, as people with coronary heart disease represent 3.5% of the total data [32].

4.2. Feature selection

To reduce the number of features entered into prediction models, binary swarm algorithms (PSO, BAT, BCS) were used in this study to find the features related to the accuracy of the prediction model. The total number of features to be reduced is 50. A method was proposed to hybrid three swarms and take into account the features selected by the three swarms to determine the features. Figure (5) shows the general structure for selecting features using swarms

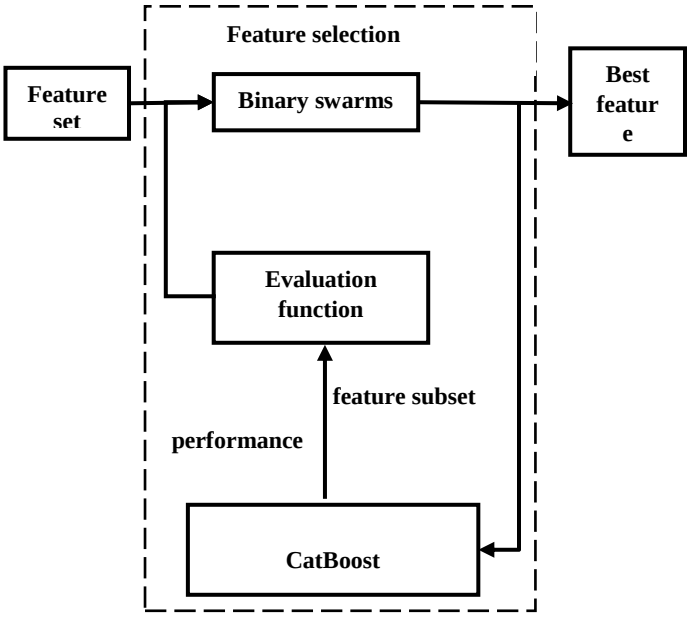

Figure 5- Structure of features selection using swarms

Selecting features begins with entering all the features and building g a swarm of particles for them. In the second stage, the selected features are evaluated by training and testing the selected model and finding the accuracy for each cycle. Depending on the model's performance, the fitness value is returned to the swarm to update the swarm parameters. This process is repeated until the required optimization is reached according to the stopping criteria. After applying the swarms, the PSO algorithm selected 8 features, while the BAT algorithm selected 30 features, and finally CS selected 29 features. Table (2) shows the parameters for each swarm.

Table 2- Parameters of PSO

Parameters	Value
Evaluate_Function	CatBoost
N	20
M	300
Minf	0
W1	0.5
c1,c2	1.1
Vmax	4

Where n is a number of population, m is Number of max iteration, minf is minimazation flag, W1 is move rate, C1, C2

are acceleration coefficients and Vmax is Limit search range

Table 3- Parameters of BAT

Parameters	Value
Evaluate Function	CatBoost
N	20
m_i	300
Minf	0
R	0.4
loud_A	0.25
Qmin	0
Qmax	2

Where n is a number of population, m_i is a number of max iteration, minf is minimazation flag, dim is number of feature, qmin is frequency minimum to step, qmax: frequency maximum to step, loud_A is value of Loudness and r is Pulse rate.

Table 4- Parameters of Binary Cuckoo Search

Parameters	Value
Evaluate Function	CatBoost
N	20
m_i	300
Minf	0
alpha and beta	0.1,1.5
Pa	0.25

Where n is a number of population, m_i is Number of max iteration, minf is minimization flag, dim is Number of feature, alpha and beta: Arguments in levy flight and pa is Probability to destroy inferior nest

4.3. Hybrid swarms feature selecting (HSFS)

To find the best features from the dataset. It is suggested to combine the outputs of the three swarms.

Binary swarms find important features according to the swarm's method, and each swarm finds a different set of features. In this paper, a hybrid between swarms was proposed based on a Merage of the outputs of the specific swarm's algorithms to find the optimal final feature subset. The vector of values for each binary swarm (binary vector) is taken, then the vector of each swarm and the rest of the swarms are combined so that the final features are taken and considered as an effective feature if they are found in any of the three swarms or were in two or all swarms

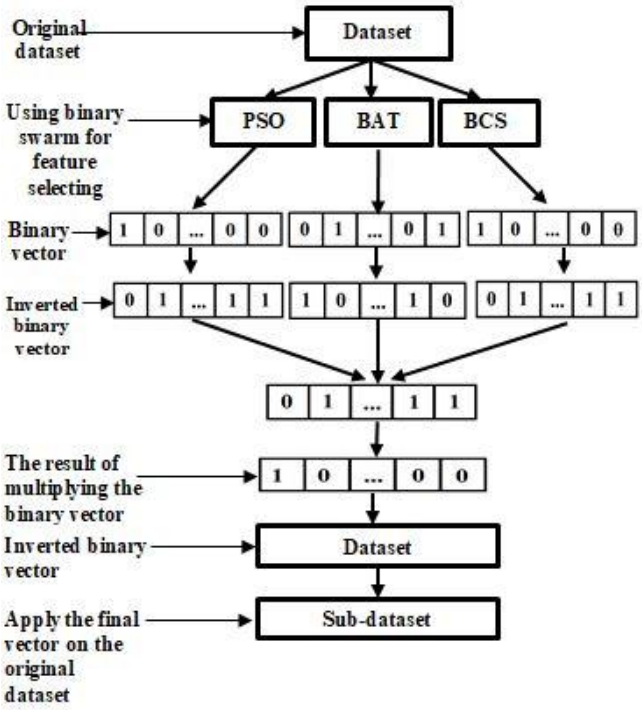

Figure 6- Framework of hybrid swarms

Specific swarms' algorithms are used to find sub-features of each swarm in the form of a one-dimensional matrix in which the number 1 represents the desired feature. In contrast, the number 0 represents the undesired feature according to the algorithm used. The 0 values are flipped to one and vice versa to use the multiplication process in the later stage. Multiplying between the inverted matrices is done to find the matrix of multiplication results; then, the final matrix is flipped to find the specific features from all the swarms. In the final stage of selecting the features, the features indicated by the value 0 are excluded from the final feature's matrix. The features indicated by the value 1 are kept finding the final features subset of the proposed method.

4.4. Training machine learning algorithms and model building

After initializing the data, analyzing it, and feature selection using single and hybrid swarms, the selected machine learning models are trained to determine the efficiency of each method to select features and to find and test the efficiency of the system as a whole. In this paper, three machine learning algorithms (naive Bayes, decision trees, CatBoost) were tested and trained using the training data set, representing 80 percent of the total data.

5. Comparing the performance of machine learning algorithms using swarms feature selection

To determine how to select the best features and compare them with the proposed hyper swarm method and the performance of each machine learning algorithm, the following is a comparison of the results of those methods and finding the best model.

5.1. Compare models using accuracy

The accuracy scale represents a basic pillar in the performance measures for machine learning algorithms [33]. The formula for accuracy is:

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

where

TP: True Positives.
TN:True Negatives.
FP: False Positives.
FN:False Negatives.

Still, it expresses the model's performance, but it does not show whether the model is biased to a certain category or the extent to which each target category is verified. Table (1) reviews the accuracy of each feature identification method and the performance of the algorithms.

Table 5- Comparing the accuracy of models concerning methods used

Methods	Accuracy
Decision Tree using Original dataset	92.30
Naive Bayes using Original dataset	54.66
CatBoost using Original dataset	95.90
Decision Tree using PSO	92.34
Naive Bayes using PSO	32.67
CatBoost using PSO	95.86
Decision Tree using BAT	91.82
Naive Bayes using BAT	9.74
CatBoost using BAT	95.84
Decision Tree using BCS	91.82
Naive Bayes using BCS	49.98
CatBoost using BCS	95.90
Decision Tree using HYBRID SWARMS	93.02
Naive Bayes using HYBRID SWARMS	54.27
CatBoost using HYBRID SWARMS	96.31

As mentioned previously, the accuracy scale does not represent the performance desired by the model. It can be seen from Table (1) that the CatBoost using the HYBRID SWARMS algorithm achieved the best performance with accuracy (96.31), followed by CatBoost using BAT and BCS accuracies have been (95.90). This indicates that classification using CatBoost is the best in most classification methods with different methods of selecting attributes. Still, this result does not express the target value of the system.

5.2. Compare models using precision

The scale of precision is the scale adopted in this paper, as it expresses the true balance value and shows the actual classification of both categories]33]. The formula for Precision is:

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

Table (5) shows the performance of models for the minority category, which is the target value.

Table 6- Comparing the precision of models to methods used for feature selecting

Methods	Precision
Decision Tree using Original dataset	0.13
Naive Bayes using Original dataset	0.08
CatBoost using Original dataset	0.45
Decision Tree using PSO	0.14
Naive Bayes using PSO	0.06
CatBoost using PSO	0.22
Decision Tree using BAT	0.1
Naive Bayes using BAT	0.04
CatBoost using BAT	0.25
Decision Tree using BCS	0.1
Naive Bayes using BCS	0.07
CatBoost using BCS	0.45
Decision Tree Using Hybrid Swarms	0.19
Naive Bayes Using Hybrid Swarms	0.08
Catboost Using Hybrid Swarms	0.56

It can be seen from the Table that the performance of the CatBoost algorithm was the best when using the method of selecting features that depend on the hybrid swarm with a precision of (0.56), which indicates an improvement in the model's performance concerning the target value (the person with a heart attack). However, this method has been the best according to the precision scale was, followed by the use of the BCS algorithm with a precision of (0.45), which indicates that the CatBoost algorithm is the best and that the hybrid swarms increased the precision of the model for the target group.

6. Conclusions

Heart attack prediction is considered one of the important topics in the health field. Building a predictive model for the classification of heart attacks faces many challenges. The most important conclusions obtained in this paper can be summarized as follows:

1- Selecting features is important to reduce the dimensions of the data set, which improves the performance of machine learning models. And give a look to health workers on the extent to which each health worker is related to the probability of having a heart attack. Swarm algorithms were used to select these features and compare them with the performance of machine learning models using the original data set and to propose a new method for selecting features (hybrid swarms). The proposed method showed an improvement in the performance of the models in terms of prediction accuracy and balancing the data.

2- Two types of machine learning algorithms were used. Single learning (DT and NB) and gradient boosting ensemble learning (CatBoost) were compared using different scales to determine the accuracy of general prediction models and the accuracy of the models for each category because the data set is unbalanced. The results showed that gradient boosting ensemble learning is better for the accuracy of the results and achieves a better balance for this type of data.

References

[1] G. Roth et al., "Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: Update From the GBD 2019 Study," Journal of the American College of Cardiology, vol. 76, pp. 2982–3021, 12/09 2020, doi: 10.1016/j.jacc.2020.11.010.

[2] Y. Yan, J.-W. Zhang, G.-Y. Zang, and J. Pu, "The primary use of artificial intelligence in cardiovascular diseases: What kind of potential role does artificial intelligence play in future medicine?," Journal of geriatric cardiology : JGC, vol. 16, pp. 585-591, 08/01 2019, doi: 10.11909/j.issn.1671-5411.2019.08.010.

[3] I. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," SN Computer Science, vol. 2, 05/01 2021, doi: 10.1007/s42979-021-00592-x.

[4] A. Iliou et al., "Metabolic phenotyping and cardiovascular disease: An overview of evidence from epidemiological settings," Heart, vol. 107, pp. heartjnl-2019, 02/19 2021, doi: 10.1136/heartjnl-2019-315615.

[5] S. Henley, R. Golden, and T. Kashner, "Statistical modeling methods: challenges and strategies," Biostatistics & Epidemiology, vol. 4, pp. 1-35, 07/22 2019, doi: 10.1080/24709360.2019.1618653.

[6] M. Kantaria, M. Buleishvili, N. Kipiani, G. Ormotsadze, and T. Sanikidze, "RISK-FACTORS OF CORONARY ARTERY DISEASE (REVIEW)," Georgian medical news, pp. 78-82, 02/01 2020.

[7] P. Dutta, S. Paul, N. Shaw, S. Sen, and M. Majumder, "Heart Disease Prediction," 2021, pp. 1-18.

[8] A. Parihar and S. Sharma, "Knowledge Discovery and Data Mining Healthcare," 2022.

[9] G. Battineni, G. Sagaro, N. Chintalapudi, and F. Amenta, "Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis," Journal of Personalized Medicine, vol. 10, 03/31 2020, doi: 10.3390/jpm10020021.

[10] M. Ghassemi, T. Naumann, P. Schulam, A. Beam, I. Chen, and R. Ranganath, "A Review of Challenges and Opportunities in Machine Learning for Health," AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science, vol. 2020, pp. 191-200, 05/30 2020.

[11] S. Dash, S. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: management, analysis and future prospects," Journal of Big Data, vol. 6, 06/19 2019, doi: 10.1186/s40537-019-0217-0.

[12] P. Kumar, S. Ambekar, M. Kumar, and S. Roy, "Analytical Statistics Techniques of Classification and Regression in Machine Learning," 2020.

[13] P. Przybyła, A. Brockmeier, and S. Ananiadou, "Quantifying risk factors in medical reports with a context-aware linear model," Journal of the American Medical Informatics Association : JAMIA, vol. 26, 03/06 2019, doi: 10.1093/jamia/ocz004.

[14] J.-K. Kim and S. Kang, "Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis," Journal of Healthcare Engineering, vol. 2017, pp. 1-13, 09/06 2017, doi: 10.1155/2017/2780501

[15] T. Kasbe and R. Pippal, Design of heart disease diagnosis system using fuzzy logic. 2017, pp. 3183-3187

[16] A. Malav, K. Kadam, and P. Kamat, "Prediction Of Heart Disease Using K-Means And Artificial Neural Network As Hybrid Approach To Improve Accuracy," International Journal of Engineering and Technology, vol. 9, pp. 3081-3085, 08/31 2017, doi: 10.21817/ijet/2017/v9i4/170904101.

[17] M. Kamboj, "Heart Disease Prediction with Machine Learning Approaches," International Journal of Science and Research (IJSR), vol. 9, no. 7, 5, July 2020 2019, doi: 10.21275/SR20724113128.

[18] Awan, Shahid & Riaz, Muhammad & Khan, Abdul. (2018). Prediction Of Heart Disease Using Artificial Neural Network. 13. 102-112. D. Shah, S. Patel, and D. Bharti, "Heart Disease Prediction using Machine Learning Techniques," SN Computer Science, vol. 1, 11/01 2020, doi: 10.1007/s42979-020-00365-y.

[19] G. Siva, S. Bindhika, M. Meghana, M. Reddy, and R. Dharmadurai, "Heart Disease Prediction Using Machine Learning Techniques," pp. 2395-0056, 10/09 2020.

[20] Walaa Adel Mahmoud a , Prof. Dr. Mohamed Aborizka a ,Prof. Dr. Fathy Ahmed Elsayed Amer2b.(2021). “Heart Disease Prediction Using Machine Learning and Data Mining Techniques: Application of Framingham Dataset”. Turkish Journal of Computer and Mathematics Education. 4864- 4870

[21] Chen, Rung & Dewi, Christine & Huang, Su & Caraka, Rezzy. (2020). “Selecting critical features for data classification based on machine learning methods”. Journal Of Big Data. 7. 26. 10.1186/s40537-020-00327-4.

[22] Wang, Dongshu & Tan, Dapei & Liu, Lei. (2018). “Particle swarm optimization algorithm: an overview”. Soft Computing. 22. 10.1007/s00500-016-2474-6.

[23] S. Akila and S. Christe, "A wrapper based binary bat algorithm with greedy crossover for attribute selection," Expert Systems with Applications, vol. 187, p. 115828, 09/01 2021, doi: 10.1016/j.eswa.2021.115828.

[24] Wang, Yechuang & Wang, Penghong & Zhang, Jiangjiang & Cui, Zhihua & Cai, Xingjuan & Zhang, Wensheng & Chen, Jinjun. (2019). A Novel Bat Algorithm with Multiple Strategies Coupling for Numerical Optimization .Journal of Mathematics. 7. 135. 10.3390/math7020135.

[25] Manar Abdulkareem Al-Abaji .(2020). “A Literature Review of Cuckoo Search Algorithm”. Journal of Education and Practice. DOI: 10.7176/JEP/11-8-01

[26] Kaviani, Pouria & Dhotre, Sunita. (2017). “Short Survey on Naive Bayes Algorithm”. International Journal of Advance Research in Computer Science and Management. 04.

[27] Kumar, Dharmender & Priyanka, N.A.. (2020). "Decision tree classifier: a detailed survey". International Journal of Information and Decision Sciences. 12. 246. 10.1504/IJIDS.2020.10029122.

[28] Reddy, V & Meghana, P & Reddy, N V Subba & Rao B, Ashwath. (2022). "Prediction on Cardiovascular disease using Decision tree and Naïve Bayes classifiers". Journal of Physics: Conference Series. 2161. 012015. 10.1088/1742-6596/2161/1/012015.

[29] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," Artificial Intelligence Review, vol. 54, 03/01 2021, doi: 10.1007/s10462-020-09896-5.

[30] Hussain, Saddam & Mustafa, Mohd & Jumani, Touqeer & Baloch, Shadi & Alotaibi, Hammad & Khan, Ilyas & Khan, Afrasyab. (2021). A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft tion. Journal of Energy Reports. 7. 4425-4436. 10.1016/j.egyr.2021.07.008

[31] Dwivedi, Ashok. (2018). "Performance evaluation of different machine learning techniques for prediction of heart disease". Journal of Neural Computing and Applications. 10.1007/s00521-016-2604-1

[32] https://catalog.data.gov/dataset/national-health-and-nutrition examination-survey-nhanes-national-cardiovascular-disease-su-00a88

[33] Vujovic, Zeljko. (2021). "Classification Model Evaluation Metrics". International Journal of Advanced Computer Science and Applications. Volume 12. 599-606. 10.14569/IJACSA.2021.0120670.

تهجين الاسراب لاختيار الميزات لنمذجة بيانات النوبات القلبية

عمر شاكر حسن [1]* ابراهيم احمد صالح [2]

omarshakir06@gmail.comi.hadedi@gmail.com

[1] علوم الحاسوب , كلية علوم الحاسوب والرياضيات , جامعة الموصل, موصل, االعراق

[2] هندسة البرمجيات, كلية علوم الحاسوب والرياضيات , جامعة الموصل, موصل, االعراق

تاريخ الاستلام:24/4/ 2022 تاريخ القبول:21/6/2022

الملخص

يعد توقع النوبات القلبية باستخدام التعلم الآلي موضوعًا مهمًا. تحتوي مجموعات البيانات الطبية سمات مختلفة منها ماهو مرتبط بالفئة المستهدفة للتنبؤ ومنها ماهو غير مرتبط بالاضافة الى ان مجموعات البيانات تكون غير متوازنة بشكل مفرط مما يؤدي الى انحياز نماذج تعلم الالة عند نمذجة النوبات القلبية. ولنمذجة مجموعة بيانات النوبات القلبية الغير متوازنة تقترح هذه الورقة تهجين الاسراب (PSO, BAT, BCS) لتحديد الميزات واعتماد الprecision لصنف الاقلية كدالة لياقة لكل سرب لتحديد الميزات ذات التاثير. ولنمذجة مجموعة البيانات التي تم تحديد السمات فيها تم اقتراح استخدام مبدا ال (boosting (Catboost كمصنف للتنبؤ بالنوبات القلبية. تمت مقارنة الطريقة المختارة لتحديد الميزات مع كل من الاسراب الثلاثة ومقارنة خوارزمية Catboost مع خوارزميات التصنيف التقليدية (naive Bayes, decision trees). توصلت الدراسة الى ان الطريقة المقترحة في دمج نتائج خورازميات الاسراب المختارة في تحديد الميزات يعد حلا واعدا في مجال تحديد الميزات ويزيد من دقة النظام وان نماذج تعلم الالة التقليدية تنحاز في حالة مجموعات البيانات الغير متوازنة و ان تحديد السمات ذات الاهمية وفقا للفئة المستهدفة ذات تاثير على اداء النماذج بالاضافة الى ان تحديد المعلمات الفائقة يقلل من انحياز النموذج المختار. حقق النموذج النهائي دقة عامة 97% بمقياس (Accuracy) و 80% بمقياس (precision) لفئة الاقلية.

الكلمات المفتاحية: التعليم الالي ، بيانات غير متوازنة، بحث الوقواق ، تحسين سرب الجسيمات ، خوارزمية الخفافيش، شجرة القرار ،خوارزمية بايز الساذجة ، خوارزمية تعزيز الفئات.