Received: October 22, 2018

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization

Youness Khourdifi[1]* **Mohamed Bahaj[1]**

[1] *Laboratory of Innovation for New Energy Technologies and Nanomaterials (LITEN), Faculty of Sciences and Techniques, Hassan 1st University, Settat, Morocco*
* Corresponding author's Email: ykhourdifi@gmail.com

Abstract: The prediction of heart disease is one of the areas where machine learning can be implemented. Optimization algorithms have the advantage of dealing with complex non-linear problems with a good flexibility and adaptability. In this paper, we exploited the Fast Correlation-Based Feature Selection (FCBF) method to filter redundant features in order to improve the quality of heart disease classification. Then, we perform a classification based on different classification algorithms such as K-Nearest Neighbour, Support Vector Machine, Naïve Bayes, Random Forest and a Multilayer Perception | Artificial Neural Network optimized by Particle Swarm Optimization (PSO) combined with Ant Colony Optimization (ACO) approaches. The proposed mixed approach is applied to heart disease dataset; the results demonstrate the efficacy and robustness of the proposed hybrid method in processing various types of data for heart disease classification. Therefore, this study examines the different machine learning algorithms and compares the results using different performance measures, i.e. accuracy, precision, recall, f1-score, etc. A maximum classification accuracy of 99.65% using the optimized model proposed by FCBF, PSO and ACO. The results show that the performance of the proposed system is superior to that of the classification technique presented above.

Keywords: Heart disease, Artificial neural network, K-nearest neighbour, Support vector machine, Naïve bayes, Random forest, Classification, Feature selection, Ant colony optimization, Particle swarm optimization, Machine learning.

1. Introduction

Cardiovascular disease (CVD) is increasing daily in this modern world. According to the World Health Organization (WHO), an estimated 17 million people die each year from cardiovascular disease, particularly heart attacks and strokes [1]. It is, therefore, necessary to record the most important symptoms and health habits that contribute to CVD. Various tests are performed prior to diagnosis of CVD, including auscultation, ECG, blood pressure, cholesterol and blood sugar. These tests are often long and long when a patient's condition may be critical and he or she must start taking medication immediately, so it becomes important to prioritize the tests [2]. Several health habits contribute to CVD. Therefore, it is also necessary to know which health habits contribute to CVD. Machine learning is now an emerging field due to the increasing amount of data. Machine learning makes it possible to acquire knowledge from a massive amount of data, which is very heavy for man and sometimes impossible [3]. The objective of this paper is to prioritize the diagnostic test and to see some of the health habits that contribute to CVD. Moreover, and above all, the different machine learning algorithms are compared using intelligent optimization algorithms. In this article, manually classified data is used. Manual classification is healthy or unhealthy. Based on a machine learning technique called classification, 70% of the data is supervised or trained and 30% is tested as part of this article.

Intelligent optimization algorithms are developed by simulating or revealing certain natural phenomena and are widely used in many research

fields because of their versatility [4, 5]. The Particle Swarm Optimization (PSO) algorithm has been successfully applied to heart disease because of its simplicity and generality [6]. However, PSO easily fell into the optimal local solution. In addition, the ACO algorithm was originally introduced for combinatorial optimization. Recently, ACO algorithms have been developed to solve continuous optimization problems. These problems are characterized by the fact that decision variables have continuous domains, unlike discrete problems [7]. Using a single optimization algorithm has the disadvantages of low accuracy and generalizability in solving complex problems. To further explore the application of intelligent optimization in bioinformatics, PSO and ACO are combined in this article, meaning that exploitation and exploration capacity are combined for binary and multi-class heart disease. In this article, the Fast Correlation-Based Feature selection (FCBF) method [8] used to remove redundant and irrelevant features, the results of the PSO optimization are considered the initial values of the ACO, and then the classification model for heart disease is constructed after the parameters are adjusted. In this study, algorithms such as K-Nearest Neighbour (K-NN), Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF) and Artificial Neural Network (ANN | MLP) are used. It can be concluded that K-Nearest Neighbour and the Random Forest are the best algorithms for the prediction and classification of heart disease dataset.

1.1 Problem statement

Previous research studies has examined the application of machine learning techniques for the prediction and classification of Heart disease. However, these studies focus on the particular impacts of specific machine learning techniques and not on the optimization of these techniques using optimised methods. In addition, few researchers attempt to use hybrid optimization methods for an optimized classification of machine learning. The most proposed studies in the literature exploit optimized techniques such as Particle Swarm Optimization and Ant Colony Optimization with a specific ML technique such as SVM, KNN or Random Forest.

In this work the Fast Correlation-Based Feature Selection (FCBF) method applied as a first step (pre-treatment). When all continuous attributes are discretized, the attribute selection attributes relevant to mining, from among all the original attributes, are selected. Feature selection, as a pre-processing step to machine learning, is effective in reducing dimensionality, eliminating irrelevant data, increasing learning accuracy and improving understanding of results. In the second step, PSO and ACO are applied to select the relevant characteristics of the data set. The best subset of characteristics selected by the characteristic selection methods improves the accuracy of the classification. Therefore, the third step applies classification methods to diagnose heart disease and measures the classification accuracy to evaluate the performance of characteristic selection methods.

The main objective of this article is the prediction heart disease using different classification algorithms such as K-Nearest Neighbour, Support Vector Machine, Naïve Bayes, Random Forest and a Multilayer Perception | Artificial Neural Network optimized by Particle Swarm Optimization (PSO) combined with Ant Colony Optimization (ACO) approaches. The weka data-mining tool is used to analyze data from a heart disease. The main contributions of this paper are:

- Extraction of classified accuracy useful for heart disease prediction
- Remove redundant and irrelevant features with Fast Correlation-Based Feature selection (FCBF) method.
- Optimizations with Particle Swarm Optimization PSO then we consider the result of PSO the initial values of Ant Colony Optimization ACO approaches.
- Comparison of different data mining algorithms on the heart disease dataset.
- Identification of the best performance-based algorithm for heart disease prediction.

1.2 Paper outline

The remainder of this paper is organized as follows. Recent work in this area is discussed in Section 2. Section 3 describes the detailed description of the proposed methodology. Section 4 explains in detail the experiments using the proposed machine learning models. Finally, Section 5 presents conclusions and future research directions.

2. Related work

Several experiments are conducted on medical data sets using multiple classifiers and features selection techniques. There is little research on the classification of the heart disease dataset. Many of them show good classification accuracy [9].

Tan et al. [10] proposed a hybrid method in which two machine learning algorithms, Support

Vector Machine (SVM) and Genetic Algorithm (G.A), are effectively combined with the wrapper approach. The LIBSVM and the WEKA data mining tool are used to analyze the results of this method. Five data sets (Iris, diabetes disease, breast cancer disease, heart disease and hepatitis) are collected from the Irvine UC machine learning repository for this experiment. After applying the hybrid GA and SVM approach, an accuracy of 84.07% is obtained for heart disease. For all diabetes data, 78.26% accuracy is achieved. The accuracy for breast cancer is 76.20%. The 86.12% accuracy is the result of hepatitis disease.

Otoom et al. [11] presented a system for analysis and follow-up. Coronary artery disease is detected and monitored by the proposed system. Cleveland Heart data are taken from the UCI. This dataset consists of 303 cases and 76 attributes/features. 13 features are used out of 76 features. Two tests with three algorithms: Bayes Naive, Support vector machine, and Functional Trees FT are performed for detection purposes. The WEKA tool is used for detection. After testing the Holdout test, the 88.3% accuracy is achieved using the SVM technique. In the cross-validation test, SVM and Bayes net provide 83.8% accuracy. The accuracy of 81.5 % is achieved after the use of FT. The 7 best features are selected using the Best First selection algorithm. For validation, cross-validation tests are used. By applying the test to the 7 best features selected, Bayes Naive achieved 84.5% accuracy, SVM provides 85.1% accuracy and FT classifies 84.5% correctly.

Parthiban et al. [12] diagnosed heart disease in diabetic patients using automatic learning methods. Naïve Bayes and SVM algorithms are applied using WEKA. A data set of 500 patients collected from the Chennai Research Institute is used. There are 142 patients with the disease and 358 patients do not have the disease. Using the Naive Bayes algorithm provides 74% accuracy. SVM provides the highest accuracy of 94.60%.

Chaurasia et al. [13] suggested using data mining approaches to detect heart disease. The WEKA data mining tool is used which contains a set of machine learning algorithms for mining purposes. Naive Bayes, J48 and bagging are used for this perspective. The UCI machine learning laboratory provides a data set on heart disease that includes 76 attributes. Only 11 attributes are used for prediction. Naive berries offer 82.31% accuracy. J48 gives 84.35% accuracy. 85.03% of the accuracy is obtained by bagging. Bagging provides a better classification rate on this data set.

Vembandasamy et al. [14] diagnosed heart disease using the Naive Bayes algorithm. Bayes' theorem is used in Naive Bayes. Therefore, Naïve Bayes has a powerful principle of independence. The data used are from one of the leading diabetes research institutes in Chennai. The data set consists of 500 patients. WEKA is used as a tool and performs classification using 70% of the Percentage Split. Naive Bayes offers 86.419% accuracy.

Some few papers proposed hybrid classification techniques.

X. Liu et al. [15] presented a study to assist in the diagnosis of heart disease using a hybrid classification system based on the ReliefF and Rough Set (RFRS) method. The proposed system consists of two subsystems: the RFRS feature selection system and a classification system with an overall classifier. A maximum classification accuracy of 92.59% was achieved according to a cross-validation scheme of the jackknife.

A. Malav et al. [16] propose an effective hybrid algorithmic approach for predicting heart disease, in order to determine and extract unknown knowledge about heart disease using the hybrid method combining the K-means clustering algorithm and the artificial neural network. The proposed model achieves an accuracy of 97%.

The common objective of all these techniques is to classify hearth disease using hybrid classification techniques. However, they used only one classification and optimization technique. The proposed approach presented a systematic way to achieve the desired results by taking into account different technical optimizations with different machine learning algorithms.

In this article, we present a hybrid approach that involves combining different techniques exploited the Fast Correlation-Based Feature Selection (FCBF) method to filter redundant features in order to improve the quality of heart disease classification. Then, we perform a classification based on different classification algorithms such as K-Nearest Neighbour, Support Vector Machine, Naïve Bayes, Random Forest and a Multilayer Perception | Artificial Neural Network optimized by Particle Swarm Optimization (PSO) combined with Ant Colony Optimization (ACO) approaches

3. Methodology

This section includes our PSO/ACO based feature selection and classification system. The main structure of the proposed system is shown in Fig. 1. Our system consists of feature selection based Fast

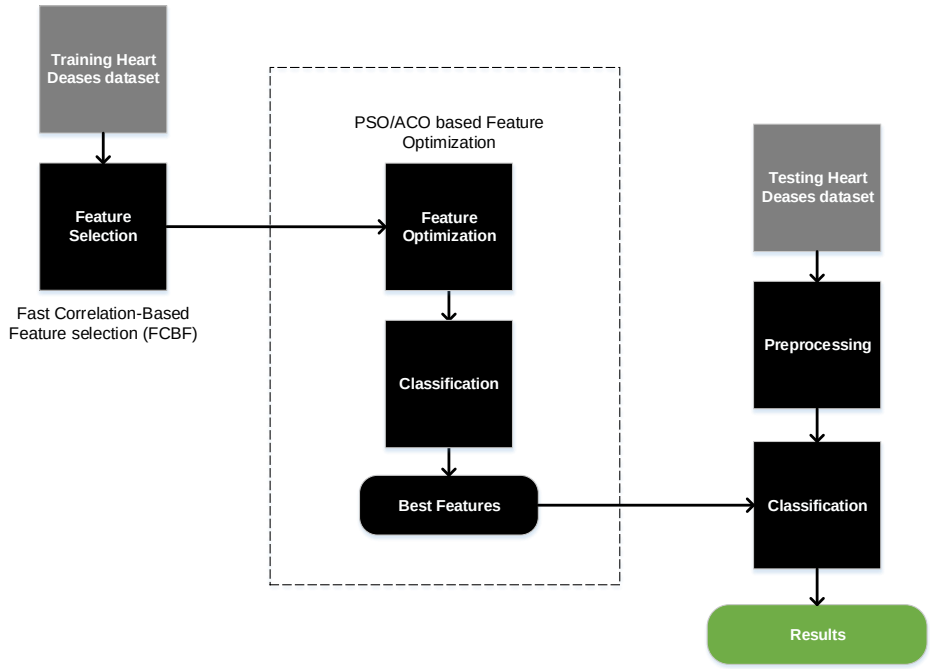

Figure. 1 The proposed architecture

Table 1. Attributes of the Heart disease dataset

Attribute	Representation	Information Attribute	Description
Age	Age	Integer	Age in years (29 to 77)
Sex	Sex	Integer	Gender instance (0 = Female, 1 = Male)
ChestPainType	Cp	Integer	Chest pain type (1: typical angina, 2: atypical angina, 3: non- anginal pain, 4: asymptomatic)
RestBloodPressure	Trestbps	Integer	Resting blood pressure in mm Hg[94, 200]
SerumCholestoral	Chol	Integer	Serum cholesterol in mg/dl[126, 564]
FastingBloodSugar	Fbs	Integer	Fasting blood sugar > 120 mg/dl (0 = False, 1= True)
ResElectrocardiographic	Restecg	Integer	Resting ECG results (0: normal, 1: ST-T wave abnormality, 2: LV hypertrophy)
MaxHeartRate	Thalach	Integer	Maximum heart rate achieved[71, 202]
ExerciseInduced	Exang	Integer	Exercise induced angina (0: No, 1: Yes)
Oldpeak	Oldpeak	Real	ST depression induced by exercise relative to rest[0.0, 62.0]
Slope	Slope	Integer	Slope of the peak exercise ST segment (1: up-sloping, 2: flat, 3: down-sloping)
MajorVessels	Ca	Integer	Number of major vessels coloured by fluoroscopy (values 0 - 3)
Thal	Thal	Integer	Defect types: value 3: normal, 6: fixed defect, 7: irreversible defect
Class	Class	Integer	Diagnosis of heart disease (1: Unhealthy, 2: Healthy)

Correlation-Based Feature selection (FCBF), feature selection based PSO (Particle Swarm Optimization) combined with ACO (Ant Colony Optimization), and classification components based on K-Nearest Neighbour, Support Vector Machine, Naïve Bayes, Random Forest and Artificial Neural Network. The training dataset is prepared according to a binary class classification problem. From the training dataset, features are selected, after that the best subset of features is optimized by our combined PSO/ACO algorithm and then, by using the selected best features, new features are classified with WEKA [17] data mining software implemented in Java. The components of our proposed system are explained in detail in the following subsections. Fig. 1 shows the proposed architecture.

3.1 Data set and attributes

The data is collected from the UCI machine learning repository. The data set is named Heart

Disease DataSet and can be found in the UCI machine learning repository. The UCI machine learning repository contains a vast and varied amount of datasets which include datasets from various domains. These data are widely used by machine learning community from novices to experts to understand data empirically. Various academic papers and researches have been conducted using this repository. This repository was created in 1987 by David Aha and fellow students at UCI Irvine. Heart disease dataset contains data from four institutions [18].

1. Cleveland Clinic Foundation.
2. Hungarian Institute of Cardiology, Budapest.
3. V.A. Medical Centre, Long Beach, CA.
4. University Hospital, Zurich, Switzerland.

For the purpose of this study, the data set provided by the Cleveland Clinic Foundation is used. This dataset was provided by Robert Detrano, M.D, Ph.D. Reason to choose this dataset is, it has less missing values and is also widely used by the research community [19].

3.2 Classification Task

From the perspective of automatic learning, heart disease detection can be seen as a classification or clustering problem. On the other hand, we formed a model on the vast set of presence and absence file data; we can reduce this problem to classification. For known families, this problem can be reduced to one classification only - having a limited set of classes, including the heart disease sample, it is easier to identify the right class, and the result would be more accurate than with clustering algorithms. In this section, the theoretical context is given on all the methods used in this research. For the purpose of comparative analysis, five Machine Learning algorithms are discussed. The different Machine Learning (ML) algorithms are K-Nearest Neighbour (KNN), Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes and Artificial Neural Network (ANN). The reason to choose these algorithms is based on their popularity [20].

3.3 Feature selection

In the heart disease datasets, the number of features can reach up to tens of thousands; the heart disease dataset has 14 attributes. Since a large number of irrelevant and redundant attributes are involved in these expression data, the heart disease classification task is made more complex. If complete data are used to perform heart disease classification, accuracy will not be as accurate, and calculation time and costs will be high. Therefore, the feature selection, as a pre-treatment step to machine learning, reduces sizing, eliminates unresolved data, increases learning accuracy, and improves understanding of results. The recent increase in the dimensionality of the data poses a serious problem to the methods of selecting characteristics with regard to efficiency and effectiveness. The FCBF's reliable method [8] is adopted to select a subset of discriminatory features prior to classification, by eliminating attributes with little or no effect, FCBF provides good performance with full consideration of feature correlation and redundancy. In this document, we first standardized the data and then selected the features by FCBF in WEKA. The number of heart disease attributes increased from 14 to 7.

3.4 Feature optimisation

3.4.1. Particle swarm optimization (PSO)

Swarm intelligence is a distributed solution to complex problems which intend to solve complicated problems by interactions between simple agents and their environment [27 - 29]. In 1995, Russel Eberhart, an electrical engineer and James Kennedy, socio-psychologist, were inspired by the living world to set up a metaheuristic: optimization by particle swarm. This method is based on the collaboration of individuals between them: each particle moves and at each iteration, the one closest to the optimum communicates its position to the others so that they can modify their trajectory. This idea is that a group of unintelligent individuals may have a complex global organization.

Due to its recent nature, a lot of research is being done on P.S.O., but the most effective so far is the extension to the framework of combinatorial optimization.

To apply the PSO it is necessary to define a research space made up of particles and an objective function to be optimized. The principle of the algorithm is to move these particles so that they find the optimum. Each of these particles is equipped with:

- From a position, i.e. its coordinates in the definition set.
- A speed that allows the particle to move. In this way, during the iterations, each particle changes position. It evolves according to its best neighbour, its best position, and its previous position. It is

this evolution that makes it possible to find an optimal particle.

- A neighbourhood, i.e. a set of particles that interact directly with the particle, especially the one with the best criterion. At any moment, each particle knows:
- It'sbest-visited position. The value of the calculated criterion and its coordinates are essentially used.
- The position of the best neighbor of the swarm that corresponds to the optimal scheduling.
- The value gives to the objective function because at each iteration it requires a comparison between the value of the criterion given by the current particle and the optimal value.

Fig. 3 shows the flowchart of the PSO algorithm. In particle swarm optimization, each individual of the population called particle. In standard PSO, after the initialization of the population, each particle update its velocity and its position in each iteration based on their own experience (pbest) and the best experience of all particles (gbest) as shown in Eql.(9 & 10). At the end of each iteration, the performance of all particles will be evaluated by predefined cost functions.

$$v^i[\mathrm{t}+1] = \mathrm{w}.v^i[t] + c_1 r_1(p^{i,best}[t] - p^i[t] + c_2 r_2(p^{g,best}[t] - p^i[t]) \quad (1)$$

$$p^i[\mathrm{t}+1] = p^i[t] + v^i[t+1] \quad (2)$$

Where, $i = 1,2,\dots,N$, N is the a number of swarm population. $v^i[t]$ is the velocity vector in $[t]th$ iteration. $p^i[t]$ represent the is the current position of the *ith* particle. $p^{i,best}[t]$ is the previous best position of *i*th particle and $p^{g,best}[t]$ is the previous best position of whole particle. To control the pressure of local and global search, *w* has been used.$c_1$and c_2 are positive acceleration coefficients which respectively called cognitive parameter and social parameter. r_1 and r_2 are random number between 0 and 1.

3.4.2. Ant colony optimization (ACO)

Ant Colony Optimization method explores to find the optimal feature subset using some iterations [30].

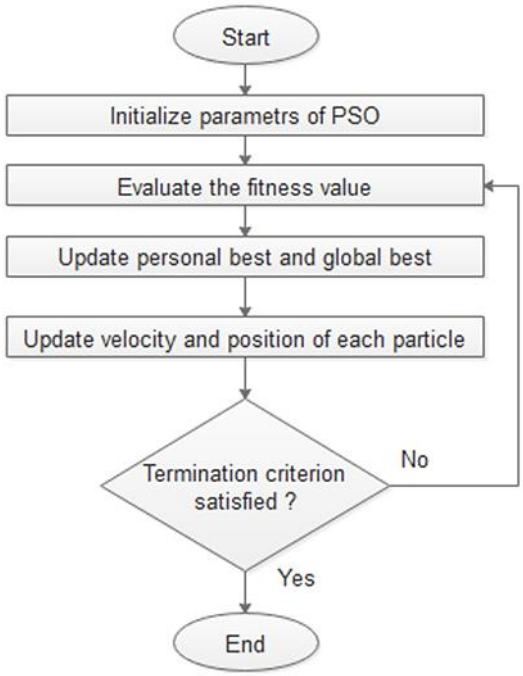

Figure. 3 The flowchart of the PSO algorithm

The main objective of the Ant Colony Optimization method is to minimize redundancy between them by selecting a subset of features. In this method, each ant in relation to the previously selected features selects the lowest similarityfeatures. Therefore, if a feature is selected by most ants, it indicates that the features has the lowest similarity with the other features. The features receive the largest amount of pheromone, and the chances of its selection by other ants will be increased in subsequent iterations. Finally, by considering the similarity between the features, the selected main features will have high pheromone values. Thus, the ACO method selects the best features with a minimum of redundancy [31].

The relevance of the features makes it possible to minimize redundancy, which will be calculated on the basis of the similarity of the features. The steps to follow to select the ACO features are described below. In this technique, before the features selection method begins, the search space must be expressed as an appropriate form for ACO. Therefore, the search space is expressed as a fully connected undirected weighted graph, $G = <F , E >$where $F = \{F1, F2 ,\dots, Fn\}$ indicates a set of all features in that each feature denotes a node in the graph, $E = \{(Fi, Fj): Fi, Fj \in F\}$ indicates the graph boundary. The connection of the boundary $(Fi, Fj) \in E$ will be set to the correlation value between *Fi and Fj*. Fig. 4 shows the illustration of the feature selection problem.

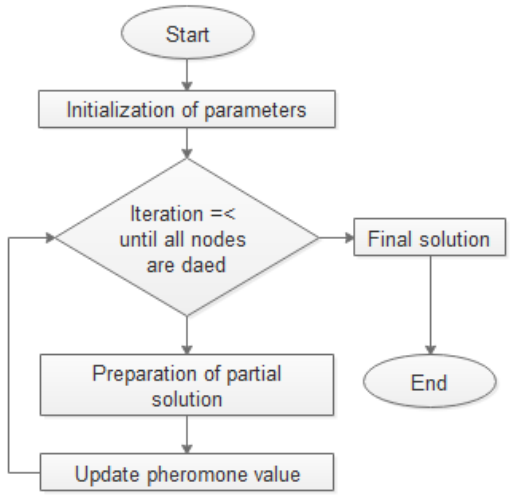

Figure. 4 The flowchart of the ACO algorithm

Table 3. Classifiers Pprformance without optimization

Evaluation criteria	K-NN	SVM	RF	NB	MLP
Time to build model (s)	0.01	0.07	0.16	0.01	0.89
Correctly classified instances	202	226	220	226	222
Incorrectly classified instance	68	44	50	44	48

Table 4. Classifiers performance optimized by FCBF

Evaluation criteria	K-NN	SVM	RF	NB	MLP
Time to build model (s)	0.01	0.09	0.55	0.01	0.58
Correctly classified instances	212	225	217	227	227
Incorrectly classified instance	58	45	53	43	43

Table 5. Classifiers performance optimized by FCBF, PSO and ACO

Evaluation criteria	K-NN	SVM	RF	NB	MLP
Time to build model (s)	0.01	0.05	0.03	0.01	0.4
Correctly classified instances	269	226	269	232	246
Incorrectly classified instance	1	44	1	38	24

4. Experiments and results

In this section, we discuss the hearth diseases datasets, experiments and the evaluation scheme. In this study, we use the Waikato Environment for Knowledge Analysis (Weka) [32].

4.1 Classification results

The aim of the entire project was to test which algorithm classifies heart disease the best with the proposed optimization methods.

The classification experiment in this paper was carried out under a Weka environment. In addition, due to the small number of selected features, 10-fold cross validation was used. For the purpose of avoiding instable operation results, each experiment was run 10 times, and the optimal classification accuracy was selected for comparison. We evaluate the effectiveness of all classifiers in terms of time to build the model, correctly classified instances, incorrectly classified instances and accuracy according to 3 steps:

1. Classifiers without optimization
2. Classifiers optimized by FCBF
3. Classifiers optimized by FCBF, PSO and ACO

4.1.1. Effectiveness

In this section, we evaluate the effectiveness of all classifiers in terms of time to build the model, correctly classified instances, incorrectly classified instances and accuracy. The results are shown in Table 3 without optimization, Table 4 optimized by FCBF and Table 5 optimized by FCBF, PSO and ACO.

In order to improve the measurement of classifier performance, the simulation error is also taken into account in this study. To do this, we evaluate the effectiveness of our classifier in terms of: Kappa as a randomly corrected measure of agreement between classifications and actual classes, Mean Absolute Error as the way in which predictions or predictions approximate possible results, Root Mean Squared Error, Relative Absolute Error, Root Relative Absolute Error, Root Relative Squared Error. The results are presented in Figs. 5, 6 and 7.

4.1.2. Accuracy results

Once the predictive model is built, we can check how efficient it is. For that, we compare the accuracy measures based on precision, recall, TP rate and FP rate values for K-NN, SVM, RF, NB and MLP. The results are shown in Table 9 without optimization, optimized by FCBF and optimized by FCBF, PSO and ACO. From the different classifiers results presented in Table 10. We can see that the best results are those generated by Classifiers optimized by FCBF, PSO and ACO. The MLP

model shows the best results in comparison with other classifiers algorithms.

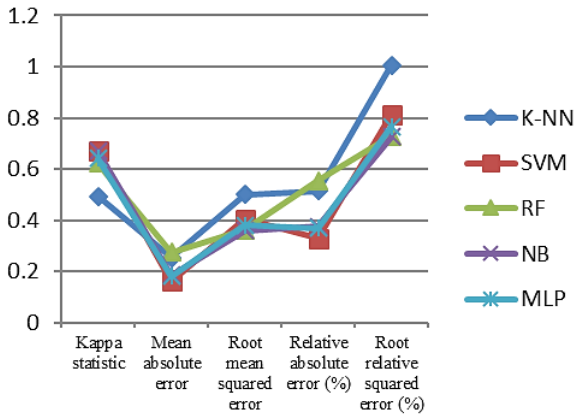

Figure. 5 Simulation error without optimization

4.1.3. Confusion matrix

Confusion matrices represent a useful way of evaluating classifier, each row of Table 10 represents rates in an actual class while each column shows predictions.

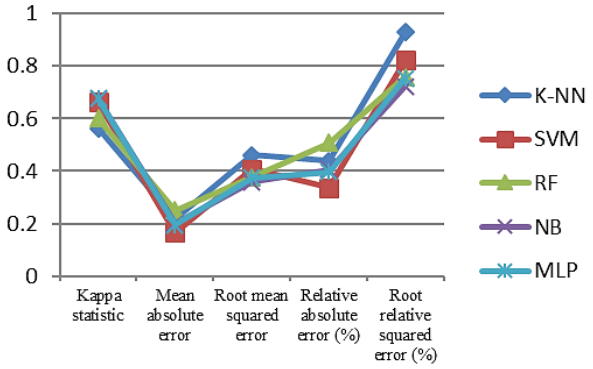

Figure. 6 Simulation error optimized by FCBF

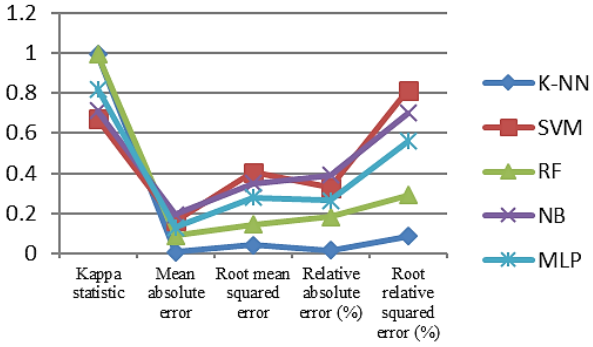

Fig. 7 Simulation error optimized by FCBF, PSO and ACO

Table 9. Accuracy / Accuracy measured by class

		TP Rate	FP Rate	Precision	Recall	F-Measure	Class
Classifiers without optimization	**K-NN**	0.753	0.258	0.785	0.753	0.769	Absence
		0.742	0.247	0.706	0.742	0.724	Presence
	SVM	0.867	0.2	0.844	0.867	0.855	Absence
		0.8	0.133	0.828	0.8	0.814	Presence
	RF	0.847	0.225	0.825	0.847	0.836	Absence
		0.775	0.153	0.802	0.775	0.788	Presence
	NB	0.867	0.2	0.844	0.867	0.855	Absence
		0.8	0.133	0.828	0.8	0.814	Presence
	MLP	0.833	0.192	0.845	0.833	0.839	Absence
		0.808	0.167	0.795	0.808	0.802	Presence
Classifiers optimized by FCBF	**K-NN**	0.833	0.275	0.791	0.833	0.812	Absence
		0.725	0.167	0.777	0.725	0.75	Presence
	SVM	0.86	0.2	0.843	0.86	0.851	Absence
		0.8	0.14	0.821	0.8	0.81	Presence
	RF	0.847	0.25	0.809	0.847	0.827	Absence
		0.75	0.153	0.796	0.75	0.773	Presence
	NB	0.873	0.2	0.845	0.873	0.859	Absence
		0.8	0.127	0.835	0.8	0.817	Presence
	MLP	0.887	0.217	0.836	0.887	0.861	Absence
		0.783	0.113	0.847	0.783	0.814	Presence
Classifiers optimized by FCBF. PSO and ACO	**K-NN**	1	0.008	0.993	1	0.997	Absence
		0.992	0	1	0.992	0.996	Presence
	SVM	0.86	0.192	0.849	0.86	0.854	Absence

	0.808	0.14	0.822	0.808	0.815	Presence
RF	0.993	0	1	0.993	0.997	Absence
	1	0.007	0.992	1	0.996	Presence
NB	0.907	0.2	0.85	0.907	0.877	Absence
	0.8	0.093	0.873	0.8	0.835	Presence
MLP	0.96	0.15	0.889	0.96	0.923	Absence
	0.85	0.04	0.944	0.85	0.895	Presence

Table 10. Confusion Matrix

		Absence	**Presence**	**Class**
Classifiers without optimization	K-NN	113	37	Absence
		31	89	Presence
	SVM	130	20	Absence
		24	96	Presence
	RF	127	23	Absence
		27	93	Presence
	NB	130	20	Absence
		24	96	Presence
	MLP	125	25	Absence
		23	97	Presence
Classifiers optimized by FCBF	K-NN	125	25	Absence
		33	87	Presence
	SVM	129	21	Absence
		24	96	Presence
	RF	127	23	Absence
		30	90	Presence
	NB	131	19	Absence
		24	96	Presence
	MLP	133	17	Absence
		26	94	Presence
Classifiers optimized by FCBF, PSO and ACO	K-NN	150	0	Absence
		1	119	Presence
	SVM	129	21	Absence
		23	97	Presence
	RF	149	1	Absence
		0	120	Presence
	NB	136	14	Absence
		24	96	Presence
	MLP	144	6	Absence
		18	102	Presence

4.2 Discussion and comparison

4.2.1. Results discussion

In this paper, we applied machine learning algorithms on heart diseasedataset to predict heart disease, based on the data of each attribute for each patient. Our goal was to compare different classification models and define the most efficient one. From all the tables above, different algorithms performed better depending upon the situation whether cross-validation, grid search, calibration and feature selection is used or not. Every algorithm has its intrinsic capacity to outperform other algorithm depending upon the situation. For example, Random Forest performs much better with a large number of datasets than when data is small while Support Vector Machine performs better with a smaller number of data sets. Performance of algorithms decreased after boosting in the data, which did not feature, selected while algorithms were performing better without boosting infeature-selected data. This shows the necessity that the data should be feature selected before applying to boost.

For the comparison of the dataset, performance metrics after feature selection, parameter tuning and calibration are used because this is a standard process of evaluating algorithms. The precision average value of the best performance without optimization it's for SVM and NB with 83.6% than RF with 81.4%. These shows SVM and NB are performing on average, after optimized by FCBF we find the best performance of precision it's for MLP with 84.2% than NB with 84% shown In Table 10. In the last stage, we compared the different algorithms with the proposed optimized model by FCBF, PSO and ACO, we find the best one is K-NN with 99.7 % than RF with 99.6 %.

4.2.2. Comparison results

We tested the proposed Classifiers optimized by FCBF, PSO and ACO against other classifications models used for hearth diseases classification described in the related work section. Table 11 compares the proposed our classification technique with previous research results. Compared to the existing methods and experiment results, we find that our optimized model performs better than the other models alone in heart disease prediction and classification. We take advantage of both the FCBF

Table 11. Performance of different methods

Model	Techniques	Disease	Tool	Accuracy
Otoom et al. [11]	Bayes Net	Heart Disease	WEKA	84.5%
	SVM			84.5%
	Functional Trees			84.5%
Vembandasamy et al. [14]	Naive Bayes	Heart Disease	WEKA	86.419%
Chaurasia et al. [13]	J48	Heart Disease	WEKA	84.35%
	Bagging	Heart Disease	WEKA	85.03%
	SVM	Heart Disease	WEKA	94.60%
Parthiban et al. [12]	Naive Bayes	Heart Disease	WEKA	74%
Tan et al. [10]	Hybrid Technique (GA + SVM)	Heart Disease	LIBSVM+WEKA	84.07%
The proposed optimized model by FCBF, PSO and ACO	K-NN	Heart Disease	WEKA	**99.65 %**
	SVM	Heart Disease	WEKA	83.55%
	RF	Heart Disease	WEKA	**99.6%**
	NB	Heart Disease	WEKA	86.15%
	MLP	Heart Disease	WEKA	**91.65%**

selection attributes and based on PSO and ACO algorithms. Thus, get higher classification accuracy than the existing models.

5. Conclusion and future work

The purpose of this work was to compare algorithms with different performance measures using machine learning. All data were pre-processed and used for test prediction. Each algorithm worked better in some situations and worse in others. K-Nearest Neighbour K-NN, and Random Forest RF and Artificial Neural Network MLP are the models likely to work best in the data set used in this study. Experimental results show that the optimization hybrid approach increase the predictive accuracy of medical data sets. The proposed methods are compared to supervised algorithms based on existing approximate sets and classification accuracy measurements are used to evaluate the performance of the proposed approaches. Therefore, the analysis section clearly demonstrated the effectiveness of hybrid PSO and ACO approaches to disease diagnosis compared to other existing approaches. The proposed optimized model by FCBF, PSO and ACO achieve an accuracy score of 99.65% with KNN and 99.6% with RF. This paper can be the first step in learning in the diagnosis of heart disease with automatic learning and it can be extended for future research. There are several limitations to this study mainly the author's knowledge base, secondly, the tools used in this study such as the processing power of the computer and thirdly the time limit available for the study. This type of study requires state-of-the-art resources and expertise in the respective fields.

Acknowledgments

I would like to thank Prof. Mohamed BAHAJ for their valuable suggestions, his relevant remarks and his perpetual advices.

References

[1] "The Atlas of Heart Disease and Stroke", [online]. http://www.who.int/cardiovascular_diseases/resources/atlas/en/

[2] J. S. Rumsfeld, K. E. Joynt, and T. M. Maddox, "Big data analytics to improve cardiovascular care: promise and challenges", *Nature Reviews Cardiology*, Vol.13, No.6, pp.350, 2016.

[3] W. Dai, T. S. Brisimi, W. G. Adams, T. Mela, V. Saligrama, and I. C. Paschalidis, "Prediction of hospitalization due to heart diseases by supervised learning methods", *International Journal of Medical Informatics*, Vol.84, No.3, pp.189–197, 2015.

[4] I. Kamkar, M. Akbarzadeh-T and M. Yaghoobi, "Intelligent water drops a new optimization algorithm for solving the Vehicle Routing Problem", In: *Proc. of IEEE International Conference on Systems, Man and Cybernetics*, pp.4142-4146, 2010.

[5] N. M. Gazzaz, M. K. Yusoff, M. F. Ramli, H. Juahir, and A. Z. Aris, "Artificial neural network modeling of the water quality index using land use areas as predictors", *Water Environment Research*, Vol.87, No.2, pp.99-112, 2015.

[6] Y. Zhang, S. Wang, and G. Ji, "A comprehensive survey on particle swarm optimization algorithm and its applications",

Mathematical Problems in Engineering, Vol.2015, Article ID 931256, 38 pages, 2015.

[7] H. M. Alshamlan, G. H. Badr and Y. A. Alohali, "Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification", *Computational Biology and Chemistry*, Vol.56, pp.49-60, 2015.

[8] L. Yu, and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution", In: *Proc. of the 20th International Conference on Machine Learning*, pp. 856-863, 2003.

[9] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications", *Journal of Intelligent Learning Systems and Applications*, Vol.9, No.01, pp.1, 2017.

[10] K. C. Tan, E. J. Teoh, Q. Yu, and K. C. Goh, "A hybrid evolutionary algorithm for attribute selection in data mining", *Expert Systems with Applications*, Vol.36, No.4, pp.8616-8630, 2009.

[11] A. F. Otoom, E. E. Abdallah, Y. Kilani, A. Kefaye, and M. Ashour, "Effective diagnosis and monitoring of heart disease", *International Journal of Software Engineering and Its Applications*, Vol.9, No.1, pp. 143-156, 2015.

[12] G. Parthiban and S. K. Srivatsa, "Applying machine learning methods in diagnosing heart disease for diabetic patients", *International Journal of Applied Information Systems*, Vol.3, No.7, pp.2249-0868, 2012.

[13] V. Chaurasia and S. Pal, "Data mining approach to detect heart diseases", *International Journal of Advanced Computer Science and Information Technology*, Vol.2, No.4, pp.56-66, 2014.

[14] K. Vembandasamy, R. Sasipriya, and E. Deepa, "Heart Diseases Detection Using Naive Bayes Algorithm", *IJISET-International Journal of Innovative Science, Engineering & Technology*, Vol.2, pp.441-444, 2015.

[15] X. Liu, X. Wang, Q. Su, M. Zhang, Y.Zhu, Q. Wang, and Q. Wang, "A hybrid classification system for heart disease diagnosis based on the rfrs method", *Computational and Mathematical Methods in Medicine*, Vol.2017, Article ID 8272091, 11 pages, 2017.

[16] A. Malav, K. Kadam, and P. Kamat, "Prediction of heart disease using k-means and artificial neural network as a hybrid approach to improve accuracy", *International Journal of Engineering and Technology*, Vol.9, No.4, 2017.

[17] S. R. W. Garner, "The Waikato environment for knowledge analysis", 2007.

[18] M. Lichman, "UCI Machine Learning Repositry", [Online]. https://archive.ics.uci.edu/, 2013.

[19] U. H. Dataset, "UCI Machine Learning Repository", [online]. https//archive.ics.uci.edu/ml/machine-learning-databases/heartdisease/ Hear.

[20] L. Van Cauwenberge, "Top 10 Machine Learning Algorithms", *Data Sci. Cent*., 2015.

[21] S. Thirumuruganathan, "A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm", [Online]. https//saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed- Introd., 2010.

[22] J. Laaksonen and E. Oja, "Classification with learning k-nearest neighbors", In: *Proc. of Neural Networks, 1996., IEEE International Conference*, Vol.3, pp.1480–1483, 1996

[23] R. Jing and Y. Zhang, "A View of Support Vector Machines Algorithm on Classification Problems", In: *Proc. of 2010 International Conference on Multimedia Communications*, pp. 13–16, 2010.

[24] G. Biau, "Analysis of a random forests model," *J. Mach. Learn. Res.*, Vol.13, pp.1063–1095, 2012.

[25] G. Louppe, "Understanding random forests: From theory to practice", arXiv Prepr. arXiv1407.7502, 2014.

[26] C. M. Bishop, "Pattern recognition and machine learning", *Inf. Sci. Stat.*, 2006.

[27] R. Poli, J. Kennedy, and T. Blackwell, "Particle swarm optimization", *Swarm Intell.*, Vol.1, No.1, pp. 33–57, 2007.

[28] J. Kennedy, "Particle Swarm Optimization", in Encyclopedia of Machine Learning (Springer), C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, pp. 760–766, 2010.

[29] Y. Shi, "Particle swarm optimization", *IEEE Connect.*, Vol.2, No.1, pp. 8–13, 2004.

[30] M. Dorigo and M. Birattari, "Ant Colony Optimization," in In Encyclopedia of machine learning (Springer), C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, pp.36–39, 2010.

[31] S. Tabakhi, P. Moradi, and F. Akhlaghian, "An unsupervised feature selection algorithm based on ant colony optimization", *Eng. Appl. Artif. Intell.*, Vol.32, pp.112–123, 2014.

[32] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," SIGKDD Explor. Newsl., Vol.11, No.1, pp.10–18, 2009.