SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES
VOL. 5, NO. 1, APRIL 2022
DOI: 10.35377/saucis.05.01.978409

Research Article

A Novel Hybrid Binary Farmland Fertility Algorithm with Naïve Bayes for Diagnosis of Heart Disease

Vafa Radpour[1], Farhad Soleimanian Gharehchopogh[2]
[1]Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran
[2]Corresponding Author; Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran; bonab.farhad@gmail.com
Received 203 August 2021; Revised 08 April 2022; Accepted 14 April 2022; Published online 30 April 2022

Abstract

One of the essential aims of intelligent algorithms concerning the diagnosis of heart disease is to achieve accurate results and discover valuable patterns. This paper proposes a new hybrid model based on Binary Farmland Fertility Algorithm (BFFA) and Naïve Bayes (NB) to diagnose heart disease. The BFFA is used for Feature Selection (FS) and the NB for data classification. FS can be employed to discover the most beneficial features. Four valid and universal UCI datasets (Hearts, Cleveland, Hungary, and Switzerland) were used to diagnose heart disease. Each dataset included 13 main features. The evaluation of the proposed model is simulated in MATLAB 2017b. The number of features in four datasets of Heart, Cleveland, Hungary, and Switzerland is equal to 13, which was reduced to six for each dataset through the BFFA to better the efficiency of the proposed model. For evaluation, the accuracy criterion, the criterion of accuracy in the proposed model for all features in the four datasets, Heart, Cleveland, Hungary, and Switzerland, is equal to 82.25%, 86.91%, 89.32%, and 89.24%, respectively. Results of the proposed model showed appropriateness in comparison to some other methods. In this paper, the proposed model was compared with other methods, and it was found that the proposed model possessed a better accuracy percentage.

Keywords: Diagnosis of Heart Disease, Binary Farmland Fertility Algorithm, Naïve Bayes Algorithm, Feature Selection, Classification, Accuracy Percentage

1. Introduction

Heart diseases are usually asymptomatic in the early stages, and heart attack and brain stroke are the first warning signs [1, 2] .Heart disease is one of the deadly silent diseases that cause more deaths than cancer in some countries. High blood pressure, high blood fats, obesity, and diabetes are risk factors for heart disease [3]. These factors gradually cause damage to the heart. Therefore, a physician must regularly monitor such risk factors for regular tests and examinations [4]. Heart disease affects the heart and blood supply to the peripheral areas of the body. Heart disease generally refers to a condition that causes the coronary arteries to narrow or become blocked, leading to heart failure, pain in the chest, or stroke [5].

Since the disease diagnosis is not easy in most cases, the physician must review and analyze the patient's tests and past decisions made for patients with similar conditions to make the right decision. In other words, the physician will need knowledge and experience. However, due to many patients and multiple tests per patient, an automated tool is required to explore patients with heart problems. One of the essential methods used for data inference is intelligent disease-diagnosis systems [6]. Collecting and recording data on heart diseases by intelligent systems is significant. Therefore, machine learning methods obtain valuable relationships between pathogens [7].

The entailing necessity in designing intelligent disease-diagnosis systems is to assist physicians in diagnosing the due disease and accelerating the testing process. Computer-based technology has significantly increased medical diagnosis, disease treatment, and patient follow-up processes [8, 9]. There is a need to diagnose it early to reduce the mortality rate related to a deficiency in coronary heart disease. As such, the existence of multiple features concerning the analysis and diagnosis of the disease proves the work of the physician brutal [10]. As a result, experts need an accurate tool that comprehensively considers these risk factors and reveals the exact impact of these ambiguous features.

In the present study, encouraged to create such a valuable and accurate tool, the researchers designed an intelligent system that identifies heart disease [11].

Concerning various interfering factors, the diagnosis of heart disease has always been a pivotal issue for physicians and teams of specialists. Recently, intelligent computational algorithms are often used in medical science to diagnose and determine the severity of various diseases. Intelligent algorithms help physicians as assisting tools in diagnosis with an incident of fewer errors. Up to now, different algorithms have been used to diagnose various diseases. In the present study, the researchers proposed a new model for the diagnosis of heart disease based on the Farmland Fertility Algorithm (FFA) [12] and NB [13]. It used the BFFA for FS and the NB algorithm for samples classification.

The FFA was presented by $Shayanfar$ and $F.S.Garahchapagh$ in 2018 [12]. The algorithm is a nature-inspired meta-heuristic algorithm that mimics fertility practices of agricultural lands. The FFA is under the inspiration of soil strategy [14, 15]. Also, the algorithm has undergone 20 mathematical tests through optimization equations. NB is among the machine learning methods for categorization [16]. As a probability hypothesis, different classes are considered separately in the technique—each new training data augmentations or diminutions the probability of correctness of the primary views. Then finally, the ideas with the highest likelihood are considered a particular class and labeled.

FS, an optimization problem, becomes a critical pre-process tool in machine learning, simultaneously minimizing feature size and maximizing model generalization. FS possesses a particular place and ranks in most research because there are various features in numerous data sets. Many are unused or do not have a positive informational aspect [17]. Non-elimination of such features does not impose a problem concerning the nature of the information. Still, they increase the computational load of the system and make it harder for accurate identification. With the increased number of features, feature space rises too. As such, data analysis and classification become significantly more complicated. Data are widely dispersed in related areas, creating significant problems concerning the supervised and unsupervised algorithms. This phenomenon is known as the size/bulk problem and is based on the fact that working with high-dimensional data is often tricky and time-consuming. A large number of features is apt to increase the noise in the data. Thus classification algorithm error rises, especially if the number of samples is smaller than the number of related features.

The significant contributions of this paper are:

- A hybrid BFFA with NB is proposed for the diagnosis of heart disease.
- A fast hybrid dimensionality reduction method for classification is proposed.
- BFFA is applied to remove the weakest feature and choose the best features.
- It is evaluated on four valid and universal UCI datasets (Heart, Cleveland, Hungary, and Switzerland).
- The proposed model shows excellent efficiency and competitive classification performance.

This paper is organized as follows: In Section 2, the related works are explained briefly. Section 3 provides a summary of the proposed model procedures. Section 4 provides experimental results and discussion. Finally, the conclusion and future work directions are summarised in Section 5.

2. Related Works

Several conducted studies in heart disease diagnosis algorithms are presented in this section. In [18], the scientists created a fuzzy algorithm for detecting heart diseases using medical knowledge and intelligence principles. Their model was based on the data set obtained from 1000 patients, including patients with heart disease and healthy individuals at $Tohid$ Hospital in $Sanandaj$ via resorting toMATLAB, featuring a fuzzy toolbox. The results of the performed experiments on the collected dataset demonstrated that the proposed model was able to 98% for accurate diagnosis of heart diseases.

In [19], the author proposed an Artificial Neural Networks (ANN) model for diagnosing heart disease. They used the data obtained from 270 patients extracted from the UCI site's valid data set, including 14 features (one feature related to class). According to the results, it was shown that the model of ANN

with multilayer perceptron structure and an accuracy of 83.33% performed the classification operation for the set of test observations. The results also indicated that the model's accuracy in classifying the heart disease response variable samples was 87.75% and 83.33% for the training and test samples.

In [20], the authors presented a model based on a hybrid Whale Optimization Algorithm (WOA) and Simulated Annealing (SA) to detect the influential factors in diagnosing heart disease. The Support Vector Machine (SVM) algorithm was used for effective disease classification. It was evaluated with the Cleveland Heart Disease Database in the UCI Valid Database. Generally, there were 13 features with 270 samples of sick and healthy individuals. The hybrid algorithm identified the number of compelling disease features based on the best solution. The proposed method identified ten valuable features with an accuracy of 87.78%.

In [21], the authors proposed a "Designing a Cardiovascular Disease Prediction System using a Support-vector Machine" model to diagnose heart disease. The dataset of 270 individuals, including 13 features, was used. The evaluation criteria in the system were classification rate and sensitivity. The system's performance was 85% and 85.8%, respectively, based on the mentioned indexes.

According to regression tree and classification, a model based on an artificial neural network (ANN) and variable selection to predict coronary artery disease was proposed [23]. The dataset included nine information variables of 13228 individuals who experienced angiography in Tehran Heart Center (4059 people without coronary artery disease and 9169 people with the disease). According to the regression tree and classification, the coronary artery disease prediction model was created based on an ANN multilayer perceptron (ANN-MLP) and variable selection method. After seven times of modeling and comparing the developed models, the final model, including all available variables occupying an area under the rocking curve of 0.754, the sensitivity of 92.41, and the accuracy of 74.19, was obtained.

[22] proposed a fuzzy-differential hybrid model based on a fuzzy expert system (FES) to predict heart disease risk. For evaluation and validation of the proposed model, a data set including 380 samples acquired from $Parsyan\ Hospital$ was used. The results showed that the FES had a functional accuracy of 85.52%, which was enhanced to 97.93% after applying the fuzzy fuzzy-genetic evolutionary model fuzzy-differential hybrid method to 97.67%. The results indicated that the proposed hybrid fuzzy genetic evolutionary model significantly improved the performance of the FES.

Sabbagh Gol [23] used the C4.5 algorithm to diagnose heart disease. The study was applied and descriptive by nature. Standard data from the credible UCI site with the Cleveland dataset was used. The database contains 303 samples (6 samples were missing). According to the utilized model, the variables of high cholesterol, gender, old age, high maximum heartbeat, and thallium scan above three and abnormal ECG had the most significant impact on coronary heart disease.

In [24] used, various ANN such as Multilayer Perceptrons (MLP), Learning Vector Quantization (LVQ), and BR to predict heart disease. The study was analytical, and its database contained 200 records of non-attributive types. The most important criteria for a disease diagnosis system were two indexes of specificity and sensitivity. These two indexes were calculated in the test and experiment phases of the study. According to the law of Back-Propagation of error, the best accuracy of the model was related to the ANN-MLP, equal to 88%. It was also observed that the elimination of discrete parameters had a positive effect on the convergence rate of the neural network and could improve the prediction accuracy by 85%. Table (1) illustrates a comparison of the proposed models for diagnosing heart disease

Table 1 Comparison of Proposed Models for Diagnosis of Heart Disease

Refs	Models	Datasets	Number of data samples	Number of features	FS	Percentage of accuracy
[18]	Fuzzy system for predicting heart disease	Collected	1000	8	8	98
[19]	Artificial Neural Networks	reference and standard	270	13	13	83.33
[20]	Hybrid WOA with SA	reference and standard	270	13	10	87.78
[21]	SVM	reference and standard	270	13	13	85
[25]	Decision tree and ANN	Collected	350	9	9	93.4
[26]	ANNs and variable selection based on regression tree	Collected	13228	9	5	74.19
[22]	Fuzzy-differential hybrid model based on a fuzzy expert system	Collected	380	5	5	97.67
[27]	C4.5	reference and standard	303	14	10	90.1
[23]	C4.5	reference and standard	297	14	5	80.2
[24]	Multi-layer artificial neural network	Collected	200	14	10	88

In [18], eight features (age, smoking, blood pressure, harmful fats, history, family, diabetes, gender) were used for evaluation. In [19], 13 features (large vessels (Nbr-ves), stress reduction (ST-dep), defect, chest pain, stress peak (Peak-ST), heart rate, angina, gender, age, static ECG (Res-elec), blood pressure (Blood-press), blood sugar, and serum cholesterol (Serum-chol)) were used for evaluation. In [20], ten features were selected as the main features for diagnosing heart disease. In [21], 13 features were chosen as the main features for diagnosing heart disease. In [25], out of 9 features, all features were added to the system for better detection. The [26] dataset includes nine risk features: age, sex, obesity, abdominal obesity, family history, smoking, high blood fats, diabetes, and high blood pressure. In [22], according to the results, the accuracy of the fuzzy expert model was 85.52%, which has increased to 97.93% after applying the hybrid Fuzzy-GA model and has grown to 97.67% after applying the hybrid Fuzzy-DE model. In [27], 14 features were used for evaluation, of which ten features had the more accurate diagnosis.

3. Proposed Model

A model based on BFFA with NB for diagnosing heart disease was proposed in the present study. MATLAB 2017b was used to simulate the proposed model, and 80% of the samples were used for the training phase and the remaining 20% for the test phase. Figure (1) shows the flowchart of the proposed model.

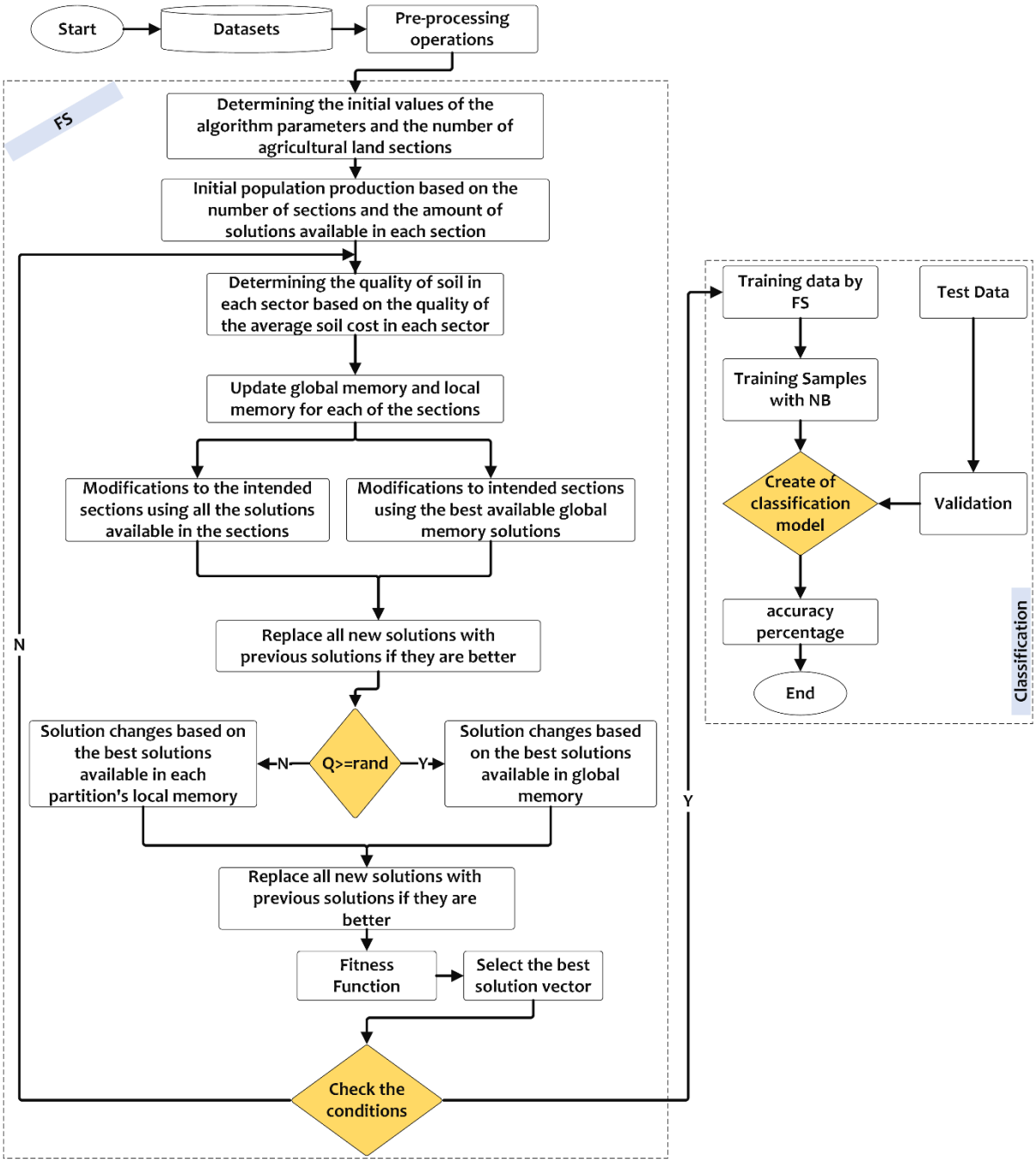

Figure 1 Proposed Model Flowcharts

3.1 Pre-processing

In this stage, the data were normalized, and the missing values were replaced. Normalization unifies and approximates the specific range of different features to some extent to achieve more accurate results. The average method was used concerning the issue of missing values. One of the most common normalization methods, according to Eq. (1), is the Min-Max method. Each data set is mapped to arbitrary intervals in this method to know the minimum and maximum values in advance. The minimum data (x_{min}) and the maximum data (x_{max}) are the minimum and maximum values in each variable (x_i) respectively.

$$N_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{1}$$

3.2 Feature Selection

In the present study, the BFFA was used for FS. Easy implementation, low complexity, and reasonable convergence rate are the main reasons for selecting the FFA regarding the FS issue. The V-Shaped function was used to convert the FFA to the binary mode. The procedure was first used by $Mirjalili$ to convert a continuous state to binary mode [28]. In constant optimization issues, the solution vectors contain valid values, and in binary optimization issues, they have values of zero or one. The steps of the BFFA are as follows.

Initial Values: The production of the initial population is defined in this step, which refers to the number of sections in the farmland and the number of solutions accessible in each part. The number of initial populations is defined according to Eq. (2).

$$\boldsymbol{N = n \times k} \tag{2}$$

In Eq. (2), The k parameter determines the number of sections connected to the optimization problem, *n* shows the number of solutions available in each land area, and parameter N indicates the total number of populations in the search space. Thus, the search space is divided into (k) parts, including several solutions.

Eq. (3) and Eq. (4) are used to determine the quality of every single section of the farmland. The quality of each section of the farmland is acquired by averaging the available solutions in each section of the farmland. Eq. (3) separates solutions in each section. As such, the average of each key is calculated separately.

$$\boldsymbol{Section_s = x(a_j), a = n * (s-1) : n * s} \\ \boldsymbol{s = \{1, 2, \dots, k\}, j = \{1, 2, \dots, D\}} \tag{3}$$

In Eq. (3), *s* shows the number of segments, x is equal to all solutions in the search space, and $j = [1, \dots, D]$ indicates the dimension of the variable x.

$$\boldsymbol{Fit\_Section_s = Mean(all\ Fit\ (x_{ij}) in\ Section_s);\ s = \{1, 2, \dots, k\}, i = \{1, 2, \dots, n\}} \tag{4}$$

In Eq. (4), $Fit\_Section$ defines the level of the quality of section's solutions of the farmland of which each part has a specific rate and indicates the average fit or suitability of all the solutions in each section in the search space. Thus, the average number of accessible solutions in each part is obtained and stored in $Fit\_Sections$ for each area of land.

Updating Memory: The local memory of each section and the global memory are updated after the solutions, and the average of each section of the farmland is determined. Some of the best states of each section and the best states of all sections are stored locally and globally.

After determining the circumstance of each part by Eq. (4), the part that has the worst condition will experience the most significant changes. According to Eq. (5) and Eq. (6), all the solutions in the worst part of the farmland would be hybrid with one in the global memory.

$$\boldsymbol{h = \alpha * rand(-1, 1)} \tag{5}$$

$$\boldsymbol{X_{new} = h * (X_{ij} - X_{MGlobal}) + X_{ij}} \tag{6}$$

In Eq. (6), $X_{MGlobal}$ is a random solution among the available solutions in the global memory, and α is a number between zero and one that must be assigned at the beginning of land fertility evaluation. The parameter X_{ij} is a solution in the worst section of the farmland picked to perform the changes, and *h* is a decimal number computed according to Eq. (5). As a result, X_{new} delivers a new solution featuring the adapted changes. After making changes in the worst section of the farmland, the other parts must be hybrids with the available solutions in the entire search space.

Soil Composition: Some available solutions in all locations are combined with the best existing solution ($Best_{Global}$) to improve the quality of the available solutions in each section. The hybridization of the supposed solution with $Best_{Global}$ or $Best_{local}$ is determined by Eq. (7).

$$\boldsymbol{H = \begin{cases} X_{new} = X_{ij} + \omega_1 * \left(X_{ij} - Best_{Global}(b)\right), Q > rand \\ X_{new} = X_{ij} + rand(0,1) * \left(X_{ij} - Best_{local}(b)\right), else \end{cases}} \tag{7}$$

In Eq. (7), a new solution is generated based on two methods. In Eq. (7), Q is a parameter in [0-1] that determines the hybridization of solutions with the best global ($Best_{Global}$). ω_1 is an integer and should be specified at the beginning of the algorithm. The value gradually declines due to the iteration of the algorithm. X_{ij} parameter is the solution selected to apply the changes from all sections. As a result, X_{new} is a new solution and is generated according to the user changes.

In the proposed model, the V-shaped function was applied to place the processes of the BFFA. Therefore, the V-shaped function continuously changes the position of the solutions in the FFA towards the binary state according to Eq. (8) [28].

$$\boldsymbol{V\left(X_i^d(t)\right) = |\left(\frac{\sqrt{\pi}}{2}\int_0^{\left(\frac{\sqrt{\pi}}{2}X_i^d(t)\right)} e^{-t^2}\,dt\right)|} \tag{8}$$

In Eq. (8) X_i^d is the continuous value of the i solution in the dth dimension in iteration *t*. The output of the V-shaped transfer function is still in a constant state between 0 and 1, so a threshold must be set to convert it to a binary value, the random point of which is determined by Eq. (9). The V-shaped function converts the solutions to binary values to FS.

$$\boldsymbol{X_i^d(t+1) = \begin{cases} 0 & if\ rand < V\left(X_i^d(t)\right) \\ 1 & if\ rand \geq V\left(X_i^d(t)\right) \end{cases}} \tag{9}$$

In Eq. (9) X_i^d shows the position of the ith solution in the population of the FFA in iteration *t* in the *d* dimension. Also, *rand* is a number between zero and one of the uniform distribution types. Therefore, the solutions of the proposed model move in a binary search space using Eq. (8).

Fit Function: The fit function is used to measure solutions. Eq. (10) is used as the fitness function for the proposed model. In Eq. (10) $\gamma_R(d)$ is classification error, R is the number of selected attributes, and the parameters α and β are in the range of zero and one, N is the total number of attributes.

$$\boldsymbol{Fitness = \alpha \times \gamma_R(d) + \beta \times \frac{|N| - |R|}{|N|}} \tag{10}$$

Ultimate Conditions: Finally, the ultimate conditions of the algorithm are examined according to iteration. If the final condition is approved, the algorithm stops. Otherwise, the algorithm continues to work to create the ultimate conditions.

3.3 Classification

NB algorithm is a probabilistic learning algorithm derived from the Bayesian theory; it is a type of classification that creates classes based on conditional probabilities. To detect the data, the NB algorithm balances the decisions of different classes and then selects the best one for them. The NB algorithm explicitly operates on various hypothetical probabilities. The NB probability theory is defined according to Eq. (11).

$$\boldsymbol{P(c|x) = \frac{P(x|c)P(c)}{P(x)}} \tag{11}$$

$$\boldsymbol{P(c|x) = P(x_1|c) \times P(x_2|c) \times \ldots \times P(x_n|c) \times P(c)} \tag{12}$$

If D is an instruction set of instances and contains class labels, then each sample is represented by an *n*-dimension feature vector $X = \{x_1, x_2, \ldots, x_n\}$, where $x1, x2, \ldots, xn$ are the attributed featured values of $A1, A2, \ldots, An$ in sample X. Assuming the existence of *m* class $C1, C2, \ldots, Cn$ and the new data sample (unlabeled), the NB classification will predict if the unknown sample X belongs to the class with the highest secondary probability.

3.4 Evaluation Criteria

Five criteria were used to evaluate the proposed model according to Table (2). True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) parameters are the main parameters concerning the evaluation criteria. TP entails the number of genuinely positive samples that have been correctly tagged. TN parameters involve the number of genuinely negative samples that have been correctly classified. FP parameters indicate the number of models without the disease but mistakenly diagnosed positive because of test results. An FN parameter shows the number of cases whose actual class is positive, and the classification algorithm has wrongly detected their category as a hostile class.

Table 2 Evaluation Criteria

Equations	**Descriptions**
$R=\frac{TP}{TP+FN}$	Recall (Sensitivity)
$P=\frac{TP}{TP+FP}$	Precision
$\frac{TN}{TN+FP}$	Specificity
$\frac{TP+TN}{TP+TN+FP+FN}$	Accuracy
$F-Measure=\frac{2*P*R}{P+R}$	F-Measure

Precision indicates the number of actual samples regarding the total number of pieces. Recall means the number of positive examples that have been correctly detected on all truly positive models. It measures how well a test has been performed in diagnosing the disease in the absence of the disease. The accuracy percentage of a classification method on the training data set is the percentage of observations of the training set correctly classified by the method used. The F-Measure criterion is based on a hybridization of accuracy and callback.

4. Evaluation and Results

The statistical population in the present study consisted of a collection of data on heart disease derived from the credible and global UCI site. The dataset had a class attribute (diagnosis) that indicated the presence or absence of heart disease according to the values of the features. The value of one distinguished affliction by the disease, and the value of two indicated non-affliction. Table (3) shows the characteristics of the datasets in detail.

Table 3 Characteristics of the Heart Disease Data set

Datasets	Main Samples	FS	Number of missing samples	Number of samples used
Heart Disease [29]	270	13	No	0
Cleveland [30]	303	13	Somewhat	297
Hungary [31]	294	13	For all samples	amendment
Switzerland [32]	123	13	For all samples	amendment

The results in Table (4) illustrate why the proposed model in this study was selected to diagnose heart disease compared to other algorithms. The number of iterations is equal to 100, and the number of initial populations is equal to 30 in all algorithms. According to the results, it is evident that the proposed model has more accuracy percentage than other algorithms. The accuracy of the proposed model on the Heart, Cleveland, Hungary, and Switzerland datasets is 88.25, 86.91, 89.32, and 89.24.

Table 4 The Results of the Models based on Different Criteria

Datasets	criteria	ABC-NB	PSO-NB	FA-NB	**Proposed Model**
Heart Disease	Accuracy	**86.13**	**87.46**	**86.92**	**88.25**
	Precision	85.73	86.52	85.76	87.42
	Recall	85.95	86.80	86.13	88.05
	F-Measure	85.84	86.66	85.94	87.73
Cleveland	Accuracy	**84.67**	**83.45**	**85.72**	**86.91**
	Precision	84.16	82.83	85.48	85.81
	Recall	84.55	83.18	85.92	86.69
	F-Measure	84.35	83.00	85.70	86.25
Hungary	Accuracy	**86.35**	**87.14**	**85.78**	**89.32**
	Precision	85.96	86.73	84.90	88.67
	Recall	86.21	87.06	85.35	89.11
	F-Measure	86.06	86.89	85.12	88.89
Switzerland	Accuracy	**87.13**	**86.49**	**85.56**	**89.24**
	Precision	86.28	85.42	84.76	88.76
	Recall	86.94	85.97	85.23	89.09
	F-Measure	86.61	85.69	84.99	88.92

Table (4) shows that in all the datasets, the proposed model carry-outs better than the other methods, demonstrating the predictive performance strength of the proposed model. Figure (2) indicates that the proposed model has a global search ability and convergence potency that outperforms other running time methods. It is evident from Table (4) and Figure (2) that the proposed model has shown competitive performance compared to FA-NB, ABC-NB, and PSO-NB.

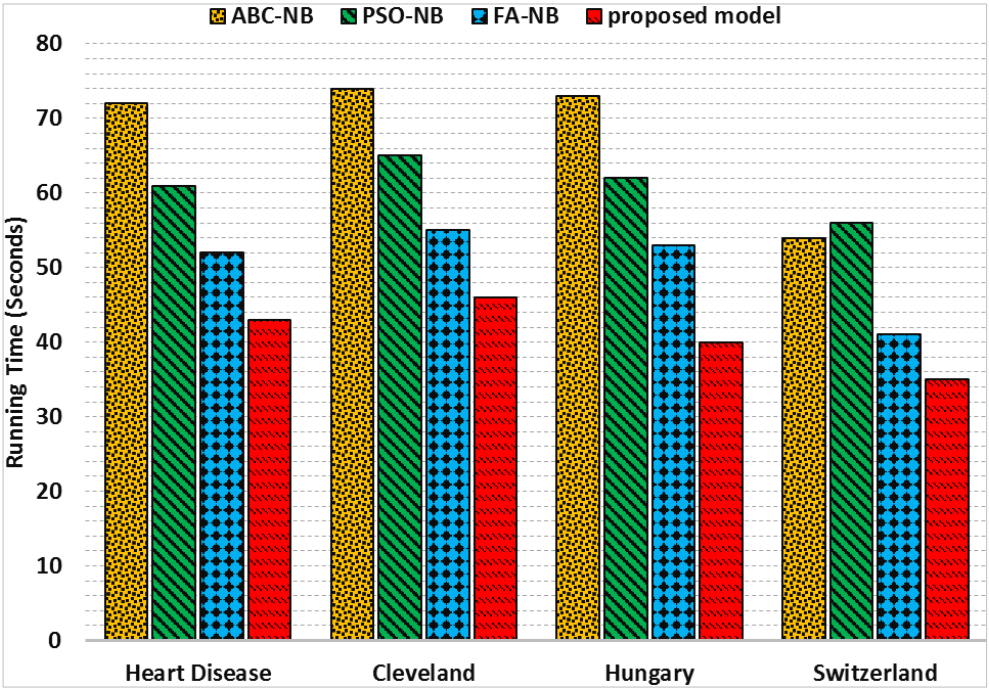

Figure 2 The running time in seconds for the proposed model and other algorithms

Figure (3) indicates convergences of ABC-NB, PSO-NB, FA-NB, and the proposed model. Additionally, Figure (3) confirms that the proposed model found the best possible optimum in iterations

71, 21, 11, and 61 for Heart Disease, Cleveland, Hungary, and Switzerland, respectively. The proposed model found the best optimum in fewer iterations than ABC-NB, PSO-NB, and FA-NB. The results revealed that the proposed model had a better convergence than ABC-NB, PSO-NB, and FA-NB.

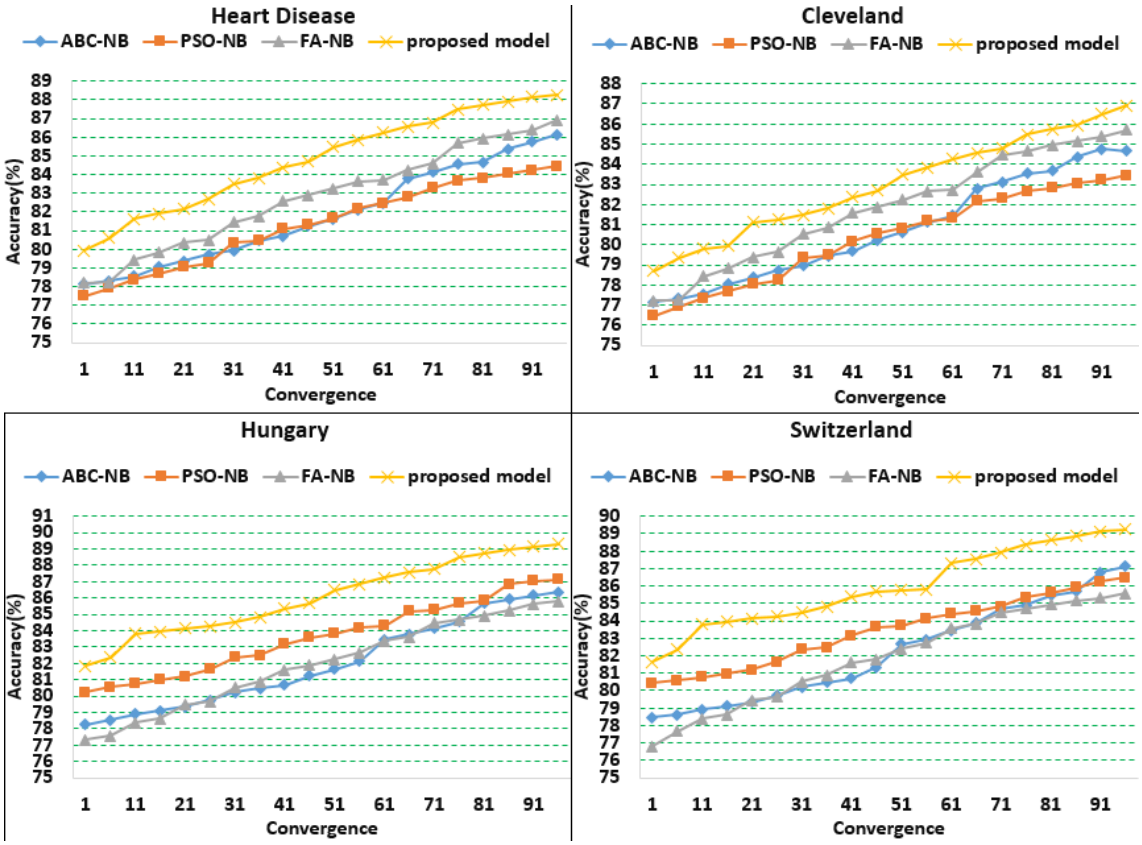

Figure 3 Convergence analysis between ABC-NB, PSO-NB, FA-NB, and proposed model

Table (5) shows the accuracy percentage of the different models based on FS. The accuracy percentage of the proposed model on the Heart data set with six features is equal to 92.23. The accuracy rate on the Cleveland dataset with six features in the proposed model is equivalent to 90.67%. The accuracy percentage of the proposed model on the Hungarian dataset with six features is equal to 92.68. The accuracy percentage of the proposed model on the Swiss dataset with six features is similar to 90.15. The outcomes indicate that the proposed model has a better accuracy percentage than other models.

Table 5 The Accuracy Percentage of the Models based on Feature Selection

Datasets	FS	ABC-NB	PSO-NB	FA-NB	Proposed Model
Heart Disease	6	91.35	91.59	91.42	92.23
	7	91.17	91.12	91.03	91.95
	9	90.07	90.25	89.79	90.56
	10	89.91	89.85	89.23	90.14
Cleveland	6	89.68	89.95	90.19	90.67
	7	89.24	89.63	89.96	90.21
	9	88.97	89.12	89.31	89.82
	10	88.65	88.84	88.91	89.28
Hungary	6	90.23	90.56	91.15	92.68
	7	89.75	90.32	90.89	92.32
	9	89.42	90.15	90.63	91.67
	10	89.12	89.92	90.25	91.38
Switzerland	6	90.72	90.25	89.91	91.15
	7	90.33	89.78	89.51	90.68
	9	89.86	89.16	88.36	90.17
	10	88.36	87.69	87.45	89.75

4.1 Comparison and Evaluation

Table (6) compares the proposed model with other models based on the accuracy percentage. The accuracy percentage of ANN, a hybridization of WOA-SA, SVM, Gravitational Search Algorithm-KNN (GSA-KNN), and Particle Swarm Optimization-KNN (PSO-KNN) on the Heart dataset was equal to 83.33, 87.78, 85, 82.96 and 83.7, respectively. The accuracy percentage of the proposed model is 88.25, which is higher than other models.

Table 6 Accuracy Percentage of Models based on Feature Selection

Datasets	Models	Refs	Accuracy
Heart Disease	ANNs	[19]	83.33
	WOA-SA	[20]	87.78
	SVM	[21]	85.00
	GSA-KNN	[33]	82.96
	PSO-KNN		83.7
	Proposed Model	-	88.25
Cleveland	C4.5	[23]	80.2
	SVM	[34]	72.36
	NB		76.19
	Deep Belief Network (DBF)		78.69
	PSO-DBF		86.44
	Proposed Model	-	86.91
Hungary	SVM	[34]	75.76
	NB		80.95
	Deep Belief Network		87.10
	PSO-DBF		87.10
	Proposed Model	-	89.32
Switzerland	SVM	[34]	75.51
	NB		76.47
	Deep Belief Network		77.78
	PSO-DBF		84.00
	Proposed Model	-	89.24

The accuracy percentage of the C4.5 [23], SVM, NB, DBF, and PSO-DBF on the Cleveland dataset were equal to 80.2, 72.36, 76.19, 78.69, and 86.44, respectively. The accuracy percentage of the proposed model was higher than other models and similar to 86.91. The SVM, NB, DBF, and PSO-DBF accuracy percentages on the Hungary dataset were 75.76, 80.95, 87.10, and 87.10. The accuracy percentage of the proposed model is 89.32, which is higher than other models.

4.2 Discussion

Analysis and results have been undertaken in (Cleveland, Hungary, and Switzerland) by using the study [34] for comparison. Accordingly, SVM, NB, Deep Belief Network, and PSO-DBF have been the existing studies to compare the proposed model. The analysis has been done concerning the accuracy obtained results in Table 6. Though the previous works applied various methods for classifying the diseases' dataset, the proposed model showed a high accuracy rate, confirming its efficiency in predicting conditions. NB showed high efficiency and excellent capability to solve complex pattern classification problems [34]. Generally, NB is a valuable and rapid method. Moreover, the selection of the most suitable neighbors by NB accelerates the diagnosis process, as it does not consider all neighbors of the evaluated item. Therefore, the proposed model is a fast and accurate decision-making system for detecting diseases.

The results indicated that the accuracy of the proposed model was 88.25% in heart, 86.91% in Cleveland, 89.32% in Hungary, and 89.24% in Switzerland. However, the proposed model indicated high accuracy of 89% than other methods. In addition, the proposed model showed high recall, precision, and f-measure rate than the other models. The proposed model showed superior performance and provided a

balance between the number of features and classification accuracy. The proposed model used the binary operation to enhance the searching process to find essential features.

5. Conclusion and Future Works

The paper processed various datasets derived from the UCI standard database. The present study predicted heart disease by applying patients with heart disease characteristics via BFFA and NB. The proposed model consisted of feature normalization, replacement of lost values, and FS. The most important features were identified in the study, and a better performance concerning precision, sensitivity, and accuracy was achieved by selecting the features based on the BFFA. The Hungary dataset's SVM, NB, DBF, and PSO-DBF accuracy were 75.51%, 76.47%, 77.87%, and 84.00%, respectively. The accuracy of the proposed model was equal to 89.24%. With a higher of 89%, this model has better performance than SVM, NB, DBF, and PSO-DBF in diagnosing heart disease. Conducted experiments and simulations showed that the medical system introduced in this study approved better performance on the heart patient database and had different accuracy percentages on different datasets. In future studies, the sensitivity of the BFFA parameters may be discovered. The BFFA may also be compared to other FS algorithms, using various large datasets and different classifiers to take better results.

References

[1] M. Langarizadeh, S. M. A. Sadr-ameli, and M. Soleymani, "Development of Vital Signs Monitoring Decision Support System for Coronary Care Unit Inpatients," Journal of Health Administration, vol. 20, no. 67, pp. 75-88, 2017.

[2] L. B. Sorkhabi, F. S. Gharehchopogh, and J. Shahamfar, "A systematic approach for pre-processing electronic health records for mining: case study of heart disease," International Journal of Data Mining and Bioinformatics, vol. 24, no. 2, pp. 97-120, 2020.

[3] M. Hassanzadeh, I. Zabbah, and K. Layeghi, "Diagnosis of Coronary Heart Disease using Mixture of Experts Method," Journal of Health and Biomedical Informatics, vol. 5, no. 2, pp. 274-285, 2015.

[4] S. M. S. Shah, F. A. Shah, S. A. Hussain, S. Batool, "Support Vector Machines-based Heart Disease Diagnosis using Feature Subset, Wrapping Selection and Extraction Methods", Computers & Electrical Engineering, vol. 84, no. 1, pp. 106628, 2020.

[5] T. Vivekanandan, and N. C. S. N. Iyengar, "Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease," Computers in Biology and Medicine, vol. 90, no. 1, pp. 125-136, 2017.

[6] S. M. S. Shaha, S. Batoolb, I. Khana, M. U. Ashrafac, S. H. Abbasa, S. A. Hussaina, "Feature extraction through parallel Probabilistic Principal Component Analysis for heart disease diagnosis," Physica A: Statistical Mechanics and its Applications, vol. 482, no. 1, pp. 796-807, 2017.

[7] S. Nazari, M. Fallah, H. Kazemipoor, A. Salehipour, "A fuzzy inference- fuzzy analytic hierarchy process-based clinical decision support system for diagnosis of heart diseases," Expert Systems with Applications, vol. 95, no.1, pp. 261-271, 2018.

[8] A.M. Alqudah, "Fuzzy expert system for coronary heart disease diagnosis in Jordan," Health and Technology, vol. 7, no. 2, pp. 215-222, 2017.

[9] S. Javadzadeh, H. Shayanfar, and F. S. Gharehchopogh, "A Hybrid Model based on Ant Lion Optimization Algorithm and K-Nearest Neighbors Algorithm to Diagnose Liver Disease," Ilam-University-of-Medical-Sciences, vol. 28, no. 5, pp. 76-89, 2020.

[10] M. H. F. Zarandi, A. Seifi, M. M. Ershadi, and H. Esmaeeli, "An Expert System Based on Fuzzy Bayesian Network for Heart Disease Diagnosis," North American Fuzzy Information Processing Society Annual Conference, NAFIPS 2017: Fuzzy Logic in Intelligent System Design, vol. 648, pp. 191-201, 2017.

[11] S. Safdar, S. Zafar, N. Zafar, and N. F. Khan, " learning based decision support systems (DSS) for heart disease diagnosis: a review," Artificial Intelligence Review, vol. 50, no. 4, pp. 597-623, 2018.

[12] H. Shayanfar, and F. S. Gharehchopogh, "Farmland fertility: A new metaheuristic algorithm for solving continuous optimization problems," Applied Soft Computing, vol. 71, pp. 728-746, 2018.

[13] Y. Jiang, H. Lin, X. Wang, and D. Lu, "A Technique for Improving the Performance of Naive Bayes Text Classification," International Conference on Web Information Systems and Mining, WISM 2011: Web Information Systems and Mining, vol. 6988, pp. 196-203, 2011.

[14] A. Benyamin, F. S. Gharehchopogh, and S. Barshandeh, "Discrete farmland fertility optimization algorithm with metropolis acceptance criterion for traveling salesman problems," International Journal of Intelligent Systems, vol. 36, no. 3, pp. 1270-1303, 2021

[15] A. Hosseinalipour, F. S. Gharehchopogh, M. Masdari, and A. Khademi, "A novel binary farmland fertility algorithm for feature selection in analysis of the text psychology," Applied Intelligence, vol. 51, pp. 4824-4859, 2021.

[16] S. Khalandi, and F. S. Gharehchopogh, "A New Approach for Text Documents Classification with Invasive Weed Optimization and Naive Bayes Classifier," Journal of Advances in Computer Engineering and Technology, vol. 4, no. 3, pp. 167-184, 2018.

[17] S. Ardam, and F. S. Gharehchopogh, "Diagnosing Liver Disease using Firefly Algorithm based on Adaboost," Journal of Health Administration, vol. 22, no. 1, pp. 61-77, 2019.

[18] V. Maihami, A. Khormehr, and E. Rahimi, "Designing an expert system for prediction of heart attack using fuzzy systems," HBI_Journals, vol. 21, no. 4, pp. 118-131, 2016.

[19] M. Kazemi, H. Mehdizadeh, and A. Shiri, "Heart disease forecast using neural network data mining technique," Ilam-University-of-Medical-Sciences, vol. 25, no. 1, pp. 20-32, 2017.

[20] Z. Hassani, and M. Khosravi, "Diagnosis of Coronary Heart Disease by Using Hybrid Intelligent Systems Based on the Whale Optimization Algorithm Simulated Annealing and Support Vector Machine," Engineering Management and Soft Computing, vol. 4, no. 2, pp. 79-93, 2019.

[21] M. S. Mahmoodi, "Designing a Heart Disease prediction System using Support Vector Machine," Journal of Health and Biomedical Informatics, vol. 4, no. 1, pp. 1-10, 2017.

[22] R. Akhoondi, and R. Hosseini, "A Novel Fuzzy-Genetic Differential Evolutionary Algorithm for Optimization of A Fuzzy Expert Systems Applied to Heart Disease Prediction," Soft Computing Journal (SCJ), vol. 6, no. 2, pp. 32-47, 2017.

[23] H. Sabbagh Gol, "Detection of Coronary Artery Disease Using C4.5 Decision Tree," Journal of Health and Biomedıcal Informatics, vol. 3, no. 4, pp. 287-299, 2017.

[24] Zabbah, M. Hassanzadeh, and Z. Koohjani, "The Effect of Continuous Parameters on The Diagnosis of Coronary Artery Disease Using Artificial Neural Networks," Journal of Torbat Heydariyeh University of Medical Sciences (Journal of Health Chimes), vol. 4, no. 4, pp. 29-39, 2017.

[25] R. Safdari, M. Ghazi Saeedi, M. Gharooni, M. Nasiri, and G. Argi, "Comparing performance of decision tree and neural network in predicting myocardial infarction," Journal of Paramedical Sciences & Rehabilitation, vol. 3, no. 2, pp. 26-35, 2014.

[26] Mahmoudi , R. A. Moghadam, M. H. Moazzam , S. Sadeghian, "Prediction model for coronary artery disease using neural networks and feature selection based on classification and regression tree," Shahrekord-University-of-Medical-Sciences, vol. 15, no. 5, pp. 47-56, 2013.

[27] H. Tahmasbi, M. Jalali, and H. Shakeri, "An Expert System for Heart Disease Diagnosis Based on Evidence Combination in Data Mining," Journal of Health and Biomedical Informatics, vol. 3, no. 4, pp. 251-258, 2017.

[28] S. Mirjalili, and A. Lewis, "S-shaped versus V-shaped transfer functions for binary Particle Swarm Optimization," Swarm and Evolutionary Computation, vol. 9, pp. 1-14, 2013.

[29] Statlog, "statlog+(heart)," 1997.[Online]. Available: https://archive.ics.uci.edu/ml/datasets/statlog+(heart). [Accessed: 25-May-2021].

[30] cleveland ," cleveland ," 2005, [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/cleveland.data. [Accessed: 25-May-2021].

[31] hungarian ," hungarian ," 1998, [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/heartdisease/processed.hungarian.data. [Accessed: 25-May-2021].

[32] switzerland ," switzerland ," 2002, [Online]. Available:https://archive.ics.uci.edu/ml/machine-learning-databases/heartdisease/processed.switzerland.data. [Accessed: 25-May-2021].

[33] S. Nagpal, S. Arora, S. Dey, A. Shreya, "Feature Selection using Gravitational Search Algorithm for Biomedical Data. Procedia Computer Science," vol. 115, no. 1, pp. 258-265, 2017.

[34] A. M. Alhassan, and W. M. W. Zainon, "Taylor Bird Swarm Algorithm Based on Deep Belief Network for Heart Disease Diagnosis," Applied Sciences, vol. 10, no. 18, pp. 1-20, 2020.