Classification Model for Heart Disease Prediction With Feature Selection Through Modified Bee Algorithm

Karunakaran Velswamy (karunasel@gmail.com)
Karunya University https://orcid.org/0000-0002-8193-4598
Rajasekar Velswamy
SRM Institute of Science and Technology
Iwin Thanakumar Joseph Swamidason
Karunya Institute of Technology and Sciences

Research Article

Keywords: Classification, Feature selection, Bee algorithm, Support Vector Machine, Navie Bayes, Decision tree and Random Forest Tree

Posted Date: May 27th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-505004/v1

Abstract

Now-a-days a healthcare field produces a huge amount of data, for processing those data some efficient techniques are required. In this paper, a classification model is developed for heart disease prediction and the attribute selection is carried out through a modified bee algorithm. The prediction of heart disease through models will help the practitioners to make a precise decision about patient health. Heart disease dataset is obtained from the UCI repository. Dataset consists of 76 features and all those seventy-six features have not contributed equal information during the time classification. In the entire attributes, some of the attributes have contributed a large amount of information at the time of classification and some of the attributes have contributed only a small amount of information during the classification task. In this paper, a modified bee algorithm is used to identify the best subset of features from the entire features in the dataset i.e., in the training phase of classification only retain those features that are contributing more information during classification and it will reduce the training time of classifiers. The experiment is analyzed with a obtained reduced subset of features by using the following classifiers such as Support Vector Machine, Navie bayes, Decision tree and Random forest. The experimental result shows that the Support Vector Machine classifier will provide a good classification accuracy, true positive rate, true negative rate, false positive rate and false negative rate compared to Navie bayes and Random forest tree classifier.

1. Introduction

Heart disease is a life-threatening disease. In the world most of the people were affected by heart diseases. Diagnosis of heart disease is also one the most important challenging task for practitioners. So, effective classification model is required to predict the heart condition of the patients, that information will help the practitioners to take precision decision in effective way in the early stage itself. Ensemble classification techniques are used for predicting heart disease. The ensemble method will improve the classification of weak algorithms through combining multiple classifiers. The experiment was carried out with the following ensemble techniques such as Boosting, Bagging, Stacking and Majority Vote. The experimental result shows that, when using boosting, stacking and bagging methods the accuracy was improved around six percentages what classification accuracy they have obtained without ensemble techniques, in majority voting method, the maximum accuracy was improved with around seven percentages. The authors concluded the majority voting method provides better classification accuracy when compared to other ensemble techniques [1]. In this paper, classification model is developed for predicting heart diseases with dimensionality reduction. Principal component analysis is used for dimensionality reduction. The method was analyzed with six classifiers such as Decision tree, Gradient Boosted tree, Logistic Regression, Multilayer Perceptron, Naïve Bayes and Random Forest. The performance of the model was evaluated by three datasets obtained from the UCI repository such as Cleveland, Hungarian and combination of Cleveland and Hungarian. The experimental result clearly shows that, the combination of Principal Component Analysis + Chi-Square + Random Forest provides better classification accuracy [2].

Khourdifi, Y., & Bahaj, M. (2019) developed an optimized classification model for predicting heart disease. Feature selection is carried out by fast correlation-based feature selection method and features were optimized by two well known approaches such as one is an Ant colony optimization and another one is Particle Swarm Optimization techniques. These hybrid approaches have been analyzed with the following classification algorithms such as K- Nearest Neighbor, Support Vector Machine, Naïve bayes, Random Forest, Multilayer Perceptron and Artificial Neural Network. The experiments were conducted with three criteria's such as 1. Classifiers without any kind of optimization, 2. Classifiers optimized with fast correlation-based feature selection method and 3. Classifiers optimized with fast correlation-based feature selection + Ant colony optimization + particle Swarm Optimization. The experimental result shows the proposed hybrid method fast correlation-based feature selection method + Ant colony optimization method + Particle swarm optimization provides better result in all aspects such as Precision, recall, F- measure and accuracy [3]. In this paper, heart disease is analyzed through two well known data mining tools such as 1. Weka and 2. Orange. The experiment was conducted with following classifiers such as 1. Naïve Bayes 2. Support Vector Machine 3. Random forest and 4. K-Nearest Neigbor. The performance was evaluated by two parameters 1. Precision and 2. Recall. The obtained result shows weka tools provides better classification accuracy than orange tool [4].

In this paper, the classification of heart disease is carried out with several data mining tools and machine learning techniques. The experiment was conducted with six data mining tools and six well known classifiers. Six data mining tools are 1. Weka 2. Orange 3. Rapid miner 4. Knime 5. Mat lab and 6. Scikit learn and six well known classifiers are 1. Support Vector Machine 2. K-Nearest Neighbor 3. Random Forest 4. Logistic Regression 5. Artificial Neural Network and 6. Naïve Bayes. The performances of the proposed system have been evaluated based on three parameters 1. Accuracy 2. Sensitivity and 3. Specificity. The experimental results concluded mat

lab data mining tool will provides better accuracy, sensitivity and specificity for all classifiers compared to other five data mining tools [5]. Huapaya et al. (2020) conducted experimental study on heart diseases detection through several supervised machine learning algorithms such as KNN, SVM, Neural Network, Random Forest and Decision tree. The performance of classifier was evaluated by three parameters 1. Precision 2. Recall and 3. F1-Score. The experimental result shows KNN classifier provides better in all aspects such as precision, recall and F1-score than other classifiers [6].

In this paper, the experiment was conducted with various machine learning methods, with feature selection techniques and without feature selection techniques for predicting heat disease. The experiment results concluded, Support Vector Machine classifier provides a better prediction rate than other classifiers without using any feature selection techniques. Naïve Bayes classifier provides a better prediction rate than other classifiers with using following feature selection methods such as Correlation based feature selection method and Fuzzy rough set. By using Fuzzy rough set and Chi square methods as a feature selection technique, Radial Basis Function Network classifier provides a better prediction rate [7]. Amin Ul Haq proposed an efficient hybrid intelligent system frame work for predicting heart disease. The hybrid intelligent system framework consists of following feature selection methods such as Relief, mRMR and LASSO and classifiers such as Logistic regression, K-NN, ANN, SVM, NB and DT. The performance of the proposed system was evaluated based on two criteria's such as with feature reduction and without features reduction. The proposed system provides good results in terms of accuracy and execution time [8].

Apurb Rajdhan et al. (2020) carried out the experimental study on heart disease prediction. The experiment was conducted with the following classifiers such as 1. Decision tree, 2. Random Forest, 3. Naïve Bayes and 4. Logistic regression. The experimental result shows that random forest classifier produces a better classification result as 90.16% as other classifiers mentioned above [9]. The following papers were conducted experimental study for predicting heart disease with various classifiers and concluded which classifier provides better prediction rate [10] [11] [12]. We tried feature selection with different optimization algorithms and we obtained the promised results [14][15][16].

The remaining of the paper is organized as follows: Sect. 2 explains outline of bee algorithm. Section 3 explains modified bees algorithm. Section 4 provides overall architecture of the proposed system. Section 5 explains feature selection using modified bee algorithm. Section 6 presents implementation details and experimental results. Section 7 presents conclusion and future work.

2. Outline Of Bee Algorithm

Many researchers are doing research on finding an efficient search algorithm to obtain a near optimal solution within a reasonable amount of execution time. Many swarm-based algorithms were developed by the researchers and all the algorithms were capable of obtaining an optimal solution within a reasonable of execution time. All the swarm-based algorithms are belonging to the category of intelligent optimization tools. Swarm based optimization are Ant Colony Optimization, Genetic algorithm, Particle Swarm Optimization algorithm, Bee algorithm, Cuttlefish optimization algorithm, Grey wolf algorithm and so on. All the above-mentioned algorithms are having capability of escaping from local optima and achieve a global optimum. In our work, we used modified bee algorithm. It is also one of the optimization algorithm it mimics the behavior of honey bees.

The characteristics of honey bees are: The scout bees randomly search the flower patches in different direction and around the more than 10 kilometers. The scout bees collected the flower patch information and return to the hive and evaluate the patches which are having good amount of nectar. Then which flower patches having good source of pollen, more number of bees were recruited for those flower patches and less number of bees recruited for those flower patches are having less nectar. How the bees will communicate to each other by means of waggle dance. The waggle dance consists of the following information such as:

1. Direction to the flower patches (It shows by the angle between sun and the flower patch)
2. Duration of the dance will show the distance between hive and flower patch.
3. Frequency of the dance will show the quality and amount of nectar.

By using the above-mentioned information, colony will recruit the bees to the flower patches in accurate way without any mistakes. Pseudo code for basic bee algorithm is shown below [13].

Pseudo code for basic Bee Algorithm

1. Initialize the population with random solutions.

2. Calculate the fitness value for each random solution.

3. While (Stopping criteria is not met) // Forming a new population through neighborhood search

4. Select the sites for neighborhood search.

5. Then recruit the bees for selected sites i.e., more bees are recruited to best e sites and the less number of bees are recruit to the remaining m-e sites and evaluate the fitness.

6. Select the fitness bee from each patch.

7. Assigning remaining bees to search randomly and evaluate their fitness.

8. End While.

The following parameter should be considered when implementing a bee algorithm.

1. In the bee algorithm 'n' stands for number of scout bees.
2. In the bee algorithm 'e' stands for elite bees.
3. In the bee algorithm 'm' stands for number of patches where selected from n visited points.
4. In the bee algorithm 'nep' stands for number of bees recruited for the patches, those patches visited by the elite bees.
5. In the bee algorithm 'nsp' stands for number of bees recruited for the remaining patches. i.e., m-e.
6. In the bee algorithm 'ngh' stands for size of the patches
7. Stopping criterion.

In the basic bee's algorithm, random initialization stage has some benefits and limitations. In this paper some modification carried out in random initialization stage through Tabu tenure value. The modified basic bee's algorithm is explained in the next section in detail.

3. Modified Bee Algorithm

In the basic bee's algorithm, random initialization stage has some advantages and disadvantages. One of the disadvantages of random initialization stage is randomly creating a 'n' number of feature subset from the entire features and there is a chance of missing one or two features were not present in any feature subset. Those features may contribute more information during classification so it will affect entire classification accuracy. To avoid this problem in our work tabu tenure value is added in the random initialization stage. Now, no features are missed out during random initialization stage. This approach will help to start the searching from the promising location and we can get better classification result what we got in basic bees algorithm.

Pseudo code for Modified Bee Algorithm

1. Initialize the population with random solutions using Tabu tenure value. (It will help us to start searching from the promising locations)

2. Calculate the fitness value for each random solution.

3. While (Stopping criteria is not met) // Forming a new population through neighborhood search

4. Select the sites for neighborhood search.

5. Then recruit the bees for selected sites i.e., more bees are recruited to best e sites and the less number of bees are recruit to the remaining m-e sites and evaluate the fitness.

6. Select the fitness bee from each patch.

7. Assigning remaining bees to search randomly and evaluate their fitness.

8. End While.

4. Overall Architecture Of The Proposed System.

The overall architecture of the proposed system is shown in Fig. 1. Input is a heart disease dataset and it consists of seventy five attributes and 303 instances\records. Proposed system consists of two stages, one is identifying the best feature subset using modified bees algorithm and in the second stage the classification task is carried out with best subset of features. The proposed system provides better true positive rate, false positive rate, true negative rate, false negative rate and accuracy what we have obtained through basic bee algorithm.

5. Feature Selection Using Modified Bees Optimization Algorithm

In this section, we will see how initialize the population with random solution though tabu tenure value. The input dataset consists of seventy five features. In original bee algorithm, first step is initializing the population with random solution. This step will not help to start searching solution from the promising locations. In our proposed system added tabu tenure value in the first step of bee algorithm that will help to start the searching solution from the promising locations. In our experiment, subset size is decided as a 13 features and maximum tabu tenure value is decided as 5 based on number of attributes present in the dataset. In our experiment tabu tenure value and feature subet size are user defined values. First, assign tabu tenure value as a zero to each feature. In first iteration randomly select 13 features from the entire features in the dataset. Those features were selected in the first iteration that tabu tenure value is replaced from 0 to 5. In the second iteration the model will select 13 features randomly, where those features are having tabu tenure value as zero. The features selected in second iteration those features tabu tenure value is replaced with five from zero and those features were selected in the first iteration, those features tabu tenure value is reduced by 1 i.e., 5 is replaced with 4. This process will continue for some k iterations. By doing this in the initialization step there is no possibility of missing any features those features are contributing moderate information during classification, definitely the searching process will start from the promising locations. Example of creating a random solution through tabu tenure value is shown Table 1. It is very difficult to explain with 75 features and subset size is 13 for easy explanation and understanding, we will explain with 18 features and features size as four with maximum tabu tenure value as 3 shown in Table 1. In the first iteration tabu tenure value is taken as zero. At the completion of first iteration, randomly the model will pick 4 features from the entire features, those features are having tabu tenure value as a zero then those features tabu tenure values is replaced with 0 from 3. In the second iteration, once again the model will randomly pick 4 features from the entire data set, those features are having tabu tenure value as zero. If the features are having tabu tenure value as other than zero the model will not pick those features upto those features tabu tenure to reach zero. In the first iteration those features are selected, those tabu tenure value is decrement by one and those features were selected in the second iteration those tabu tenure value is replaced from 0 to 3. This process will continue for k iteration. k is an user defined value. The example of initializing the population with random random solution through tabu tenure value is shown in Table 1.

Table 1
Initialize the population with random solution through tabu tenure value.

Features	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18
Tabu tenure value	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Completion of iteration 1(Selected features are: F2, F8, F11,F13) and updated tenure value.	0	3	0	0	0	0	0	3	0	0	3	0	3	0	0	0	0	0
Completion of iteration 2(Selected features are: F1, F4, F9,F14) and updated tenure value.	3	2	0	3	0	0	0	2	3	0	2	0	2	3	0	0	0	0
Completion of iteration 3(Selected features are: F3, F7, F17, F18) and updated tenure value.	2	1	3	2	0	0	3	1	2	0	1	0	1	2	0	0	3	3
Completion of iteration 4(Selected features are: F5, F12, F15, F16) and updated tenure value.	1	0	2	1	3	0	2	0	1	0	0	3	0	1	3	3	2	2
Completion of iteration 4(Selected features are: F2, F13, F6, F8)	0	3	1	0	2	3	1	3	0	0	0	2	3	0	2	2	1	1
: :																		
Completion of K iteration (Selected features are F5, f10, F12, f16)	1	0	2	0	3	0	1	2	2	3	1	3	2	1	0	3	0	0

6. Implementation Details And Experimental Results

The experiments were carried out with the following three criteria's such as 1. Experiment is conducted with entire dataset without any data reduction techniques. 2. Experiment is conducted with reduced dataset (Bee algorithm) and 3. Experiment is conducted with reduced dataset (Modified Bee algorithm). The experiments were conducted with following configuration system and software, i5 processor with 3 GB ram and the model is implemented by using Matlab 9.2. The experiment is conducted with k cross validation. The performance is calculated average of 'k' iterations. The performance of the model is evaluated by the following parameters for all three criteria's such as false positive rate, false negative rate, true positive rate, true negative rate and accuracy. The formula for calculating true positive rate, false positive rate, false negative rate, true negative rate and accuracy are shown below.

True Positive Rate = Number of heart disease is correctly classified as a heart disease/Number of heart disease is present in test dataset * 100. ……. (1)

False Positive Rate = Number of normal data incorrectly classified as heart disease/Total number of normal data present in the test dataset *100…… (2)

False Negative Rate = Number of heart disease data incorrectly classified as normal data/Number of heart disease present in dataset*100…… (3)

True Negative Rate = Number of normal data is correctly classified as normal/Total number of normal data present data in test dataset* 100…….. (4)

Accuracy = True positive + True negative / True positive + True Negative + False positive + False Negative *100…….. (5)

In our work, fitness function is taken as maximum of accuracy. The experiment result is shown in Table 2 and Table 3 and graphical representation is shown in Fig. 2 to 4.

Table 2
True Positive Rate, False Positive Rate, False Negative Rate and True Negative Rate using SVM, Navie bayes and Random Forest Tree classifiers for heart disease dataset.

Heart disease Dataset		Number of features	Number of instances	True Positive Rate	False Positive Rate	False Negative Rate	True Negative Rate
Support Vector Machine	Entire dataset (criteria 1)	75	303	84.12 ± 0.26	15.32 ± 0.21	14.37 ± 0.56	85.63 ± 0.87
	Reduced dataset (basic Bee algorithm) (Criteria 2)			81.16 ± 0.34	16.46 ± 0.37	14.27 ± 0.64	85.83 ± 0.22
	Reduced dataset (Modified Bee algorithm) (Criteria 3)			84.98 ± 0.87	15.11 ± 0.23	13.03 ± 0.31	86.97 ± 0.88
Navie bayes	Entire dataset (criteria 1)			80.57 ± 0.34	16.98 ± 0.19	15.81 ± 0.46	82.9 ± 0.74
	Reduced dataset (basic Bee algorithm) (Criteria 2)			79.11 ± 0.39	16.56 ± 0.85	16.18 ± 0.63	80.19 ± 0.72
	Reduced dataset (Modified Bee algorithm) (Criteria 3)			82.64 ± 0.54	15.56 ± 0.84	15.44 ± 0.31	83.34 ± 0.67
Random Forest Tree	Entire dataset (criteria 1)			82.89 ± 0.52	15.76 ± 0.21	15.91 ± 0.91	82.12 ± 0.33
	Reduced dataset (basic Bee algorithm) (Criteria 2)			83.33 ± 0.45	15.66 ± 0.34	14.84 ± 0.72	84.82 ± 0.24
	Reduced dataset (Modified Bee algorithm) (Criteria 3)			83.11 ± 0.22	15.34 ± 0.56	14.17 ± 0.73	84.34 ± 0.45

Table 3
Accuracy rate of heart disease dataset using SVM, Naïve Bayes and Random Forest Tree.

Classifiers		Number of features	Number of instances	Accuracy rate
Support Vector Machine	Entire dataset (criteria 1)	75	303	85.35 ± 0.26
	Reduced dataset (basic Bee algorithm) (Criteria 2)			84.69 ± 0.37
	Reduced dataset (Modified Bee algorithm) (Criteria 3)			85.85 ± 0.65
Naïve Bayes	Entire dataset (criteria 1)			83.93 ± 0.27
	Reduced dataset (basic Bee algorithm) (Criteria 2)			83.24 ± 0.67
	Reduced dataset (Modified Bee algorithm) (Criteria 3)			84.61 ± 0.45
Random Forest Tree	Entire dataset (criteria 1)			84.53 ± 0.83
	Reduced dataset (basic Bee algorithm) (Criteria 2)			85.20 ± 0.43
	Reduced dataset (Modified Bee algorithm) (Criteria 3)			85.20 ± 0.26

The experiment results clearly shows that SVM classifer provides little bit better in terms of accuracy on criteria 1 and criteria 3 when comapred to the naïve bayes and Random forest tree classifiers. Random forest tree classifiers provides little bit better in terms of accuracy on criteria 2. The experiment results is shown is Table 3.

7. Conclusion And Future Work

Now-a-days a health care field produces a huge amount of data, for processing those data some efficient techniques are required. In this work, classification model is developed for heart disease prediction and the attribute selection is carried out through modified bee algorithm. The experiment results clearly shows that SVM classifer provides little bit better in terms of accuracy on criteria 1 and criteria 3 when comapred to the naïve bayes and Random forest tree classifiers. Random forest tree classifiers provides little bit better in terms of accuracy on criteria 2. The proposed system provides almost more or less similar accuracy rate when compared to the reduced dataset through basic bee algorithm and with entire dataset. In future the experiement may tried with large datasets.

Declarations

Compliance with Ethical Standards statements:

i. **Ethical approval**- The research was not carried out with any human beings, the research work was carried out with only available datasets, and those datasets were available in UCI repository and Kaggle repository. (Ethical Approval is not required)

ii. **Funding details (In case of Funding)-** Fund is not received from any organization (Nil).

iii. **Conflict of interest**-Nil (There is no conflict of interest in this research).

iv. **Informed Consent**-In this research, we took only the dataset available in the UCI repository and Kaggle repository, we didn't take any informed consent from any organization.

Authorship contributions:

Author1: V. Karunakaran

i. Design and Implementation tools

ii. Data Collections
iii. Determining the performance metrics
iv. Implementation

Author2: V. Rajasekar

i. Collections of existing papers
ii. Carried out the details literature survey
iii. Carried out the performance analysis.

Author 3: S. Iwin

i. Carried out the documentation
ii. Grammar checking

References

1. Latha CBC, Jeeva SC (2019) Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked 16:100203
2. Garate-Escamilla AK, Hassani AHE, Andres E (2020) Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked, 100330
3. Khourdifi Y, Bahaj M (2019) Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. Int J Intell Eng Syst 12(1):242–252
4. Sarangam Kodati DRV (2018) Analysis of heart disease using in data mining tools Orange and Weka. Global Journal of Computer Science and Technology
5. Tougui I, Jilbab A, El Mhamdi J. Heart disease classification using data mining tools and machine learning techniques. Health Technol. (2020). https://doi.org/10.1007/s12553-020-00438-1
6. Huapaya HD, Rodriguez C, Esenarro D (2020) Comparative analysis of supervised machine learning algorithms for heart disease detection
7. Cengiz, Gazeloğlu (2020) Prediction of heart disease by classifying with feature selection and machine learning methods. Progress in Nutrition 22:N. 2: 660–670. DOI:10.23751/pn.v22i2.9830
8. Haq AU, Li JP, Memon MH, Nazir S, Sun R (2018) A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems, 2018
9. Rajdhan A, Agarwal A, Sai M, Ravi D, Ghuli P Heart Disease Prediction using Machine Learning, International Journal of Research and Technology, Vol. 9 Issue 04, April-2020. PP. 659–662
10. Alotaibi FS (2019) Implementation of machine learning model to predict heart failure disease. (IJACSA) International Journal of Advanced Computer Science and Applications, 10(6)
11. Singh AK, A Comparative Study on Disease Classification using Machine Learning Algorithms, 2nd International Conference on Advanced Computing and Software Engineering (ICACSE-2019)
12. Rajesh N, Maneesha T, Ahaik, Hafeez (2018) Hari Krishna, Prediction of Heart Disease using Machine Learning Algorithms. International Journal of Engineering Technology 7(2):363–366
13. Pham DT, Ghanbarzadeh A, Koc E, Otri S, Rahim S, Zaidi M (2005) The Bees Algorithm. Technical Note. Manufacturing Engineering Centre, Cardiff University, UK
14. Suganthi M, Karunakaran V (2019) Instance selection and feature extraction using cuttlefish optimization algorithm and principal component analysis using decision tree. Cluster Computing 22(1):89–101
15. Karunakaran V, Suganthi M, Rajasekar V (2020) Feature selection and instance selection using cuttlefish optimisation algorithm through tabu search. International Journal of Enterprise Network Management 11(1):32–64
16. Karunakaran V, Joseph SI, Teja R, Suganthi M, Rajasekar V (2019) A wrapper based feature selection approach using bees algorithm for extreme rainfall prediction via weather pattern recognition through svm classifier. International Journal of Civil

Figures

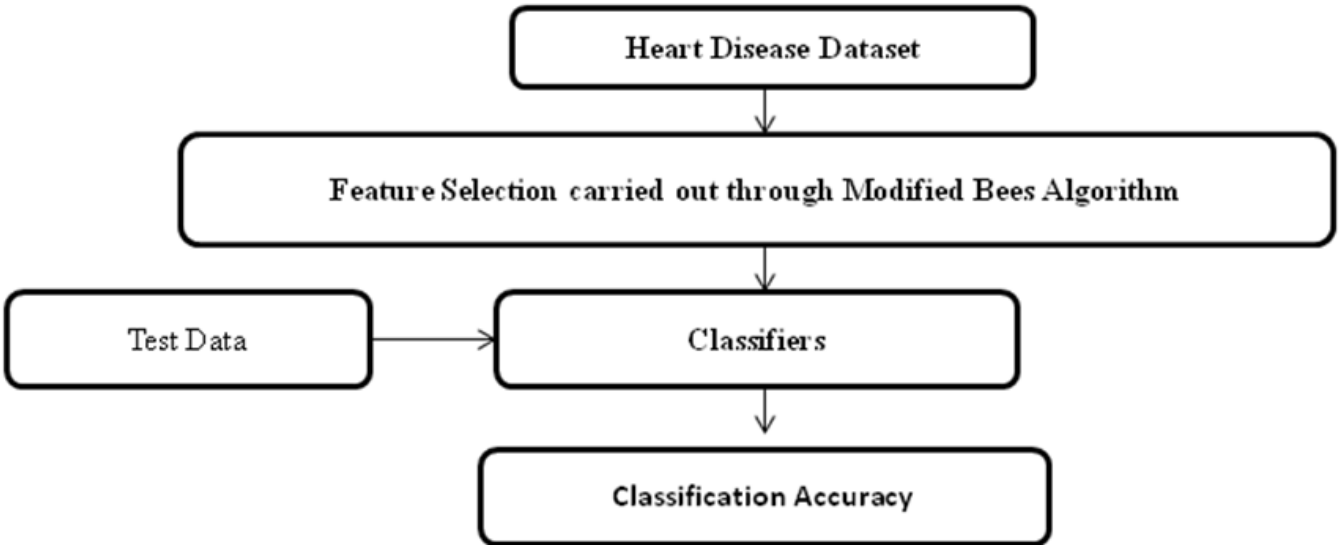

Figure 1

Overall architecture of the proposed system

Figure 2

TPR,FPR FNR and TNR using SVM classifier.

Figure 3

TPR,FPR FNR and TNR using Naïve Bayes classifier.

Figure 4

TPR,FPR FNR and TNR using Naïve Bayes classifier.

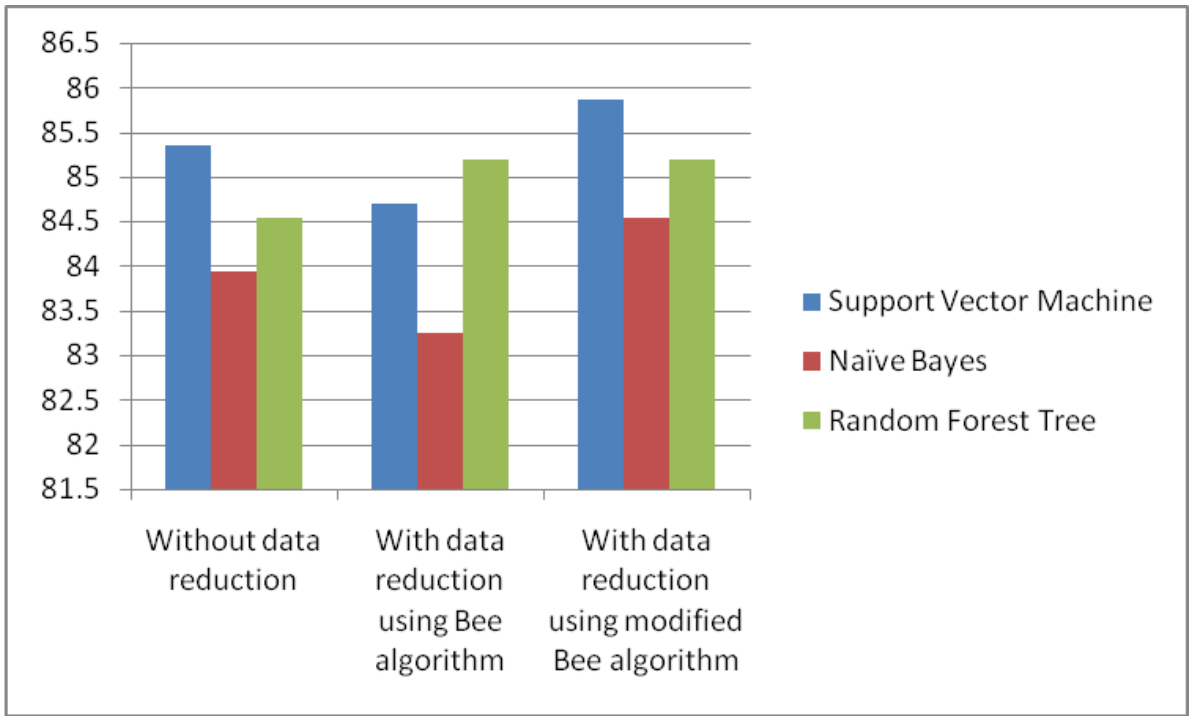

Figure 5

Accuracy of three criteria's using SVM, Naïve Bayes and Random Forest Tree.