RESEARCH

A hybrid cost-sensitive ensemble for heart disease prediction

Zhenya Qi[1] and Zuoru Zhang[2*]

*Correspondence:
zhangzuoru@tju.edu.cn
[2]School of Mathematical Science, Hebei Normal University, Yuhua District, 050024 Shijiazhuang, PR China
Full list of author information is available at the end of the article

Abstract

Heart disease is the primary cause of morbidity and mortality in the world. It includes numerous problems and symptoms. The diagnosis of heart disease is difficult because there are too many factors to analyze. What's more, the misclassification cost could be very high. In this paper, we firstly propose a cost-sensitive ensemble model to improve the efficiency of diagnosis and reduce the misclassification cost. The proposed model contains five heterogeneous classifiers: random forest, logistic regression, support vector machine, extreme learning machine and k-nearest neighbor. Then, experiments are done on three datasets from UCI machine learning repository. The best performance is achieved by the proposed model according to ten-fold cross validation. The statistical tests demonstrate that the performance of the proposed model is significantly superior to individual classifiers, and the efficiency of classification is distinctively improved by Relief algorithm.

Keywords: cost-sensitive; ensemble; heart disease

1 Introduction

Heart disease is any disorder that influences the heart's ability to function normally [1]. As the leading cause of death, heart disease is responsible for nearly 30% of the global deaths annually [2]. In China, it is estimated that 290 millon people are suffering from heart disease, and the rate of death caused by heart disease is more than 40% [3]. According to The European Society of Cardiology (ESC), nearly half of the heart disease patients die within initial two years [4]. Therefore, accurate diagnosis of heart disease in early stages is of great importance in improving security of heart [5].

However, as it's associated with numerous symptoms and various pathologic features such as diabetes, smoking and high blood pressure, the diagnosis of heart disease remains a huge problem for less experienced physicians [6]. In order to detect heart disease, several diagnostic methods have been developed, Coronary angiography (CA) and Electrocardiography (ECG) are the most widely used among them, but they both have serious defects. ECG may fail to detect the symptoms of heart disease in its record [7] while CA is invasive, costly and needs highly-trained operators [8].

Computer-aided diagnostic methods based on machine learning predictive models are noninvasive and provide proper and objective diagnoses, and hence can reduce the suffering of patients [9]. Various machine learning predictive models [10–14] have been developed and widely used as classifiers to assist doctors in diagnosing heart

disease. Dogan et al.[15] built a RF classification model for symptomatic heart disease. The clinical characteristics of the 1545 and 142 subjects were used for training and testing respectively, and the classification accuracy was 78%. Detrano et al. [16] proposed a LR classifier for heart disease classification and obtained an accuracy of 77% in 3 patient test groups. Gokulnath and Shantharajah [17] proposed a classification model based on genetic algorithm (GA) and SVM, obtaining an accuracy of 88.34% on Cleveland heart disease dataset. Subbulakshmi et al. [18] performed a detailed analysis of different activation functions of ELM using Statlog heart disease dataset. The results indicated that ELM achieved an accuracy of 87.5%, higher than other methods. Duch et al. [19] used KNN classifier to predict heart disease on Cleveland heart disease dataset and achieved an accuracy of 85.6%, superior to other machine learning techniques.

It is realized that no single model exists that is superior for all classification problems, because different machine learning algorithms consider datasets with different features in different aspects [20]. One way to overcome the limitations of a single classifier is to use an ensemble model. An ensemble model is the combination of multiple sets of classifiers, it outperforms the individual classifiers because the variance of error estimation is reduced [21]. In recent years, many ensemble approaches have been proposed to improve the performance of heart disease diagnosis systems. For instance, Das et al. [22] proposed a neural networks ensemble and obtained 89.01% classification accuracy from the experiments made on the data taken from Cleveland heart disease dataset. Bashir et al. [23] employed the ensemble of five heterogeneous classifiers on five heart disease datasets. The proposed ensemble classifier achieved the high diagnosis accuracy of 87.37%. Khened et al. [24] presented an ensemble system based on deep fully convolutional neural network (FCN) and achieved a maximum classification accuracy of 100% on Automated Cardiac Diagnosis Challenge (ACDC-2017) dataset. Therefore, we use an ensemble classifier to predict the presence or absence of heart disease in present study.

From the previous studies, it is observed that traditional medical decision support systems usually focused only on the maximization of classification accuracy without taking the unequal misclassification costs between different categories into consideration. However, in the field of medical decision making, it is often the minority class that is of higher importance [25]. Further, the cost associated with missing a patient (false negative) is much higher than that of mislabeling a healthy instance (false positive) [26]. Therefore, traditional classifiers inevitably result in a defective decision support system. In order to overcome this limitation, in this paper we combine the classification results of individual classifiers in a cost-sensitive way so that classifiers that help reduce the costs gain more weights in the final decision.

The rest of the paper is organized as follows. Section 2 offers brief background information concerning Relief algorithm and each individual classifier. Section 3 presents the framework of the proposed cost-sensitive ensemble diagnosis model. Section 4 illustrates the research design of this paper in detail. Section 5 describes the experimental results and compares the ensemble model with individual classifiers and previous methods. Finally, the conclusions and directions for future works are summarized in Section 6.

2 Backgrounds and Preliminaries

2.1 Relief Feature Selection Algorithm

Relief is a kind of famous filter feature selection algorithm which adopts a relevant statistics to measure the importance of the feature. This statistics can be seen as the weight of each feature. Top k features of bigger weights are selected. Therefore, the key is to determine the relevant statistics [27].

Assume $D = \{(x_1, y_1), (x_2, y_2), ...(x_m, y_m)\}$ is a dataset. x_i is an input feature vector and y_i is a class label corresponding to x_i. First, select a sample x_i randomly. Then, Relief attempts to find out its nearest sample $x_{i,nh}$ from samples of its same class and nearest sample $x_{i,nm}$ from samples of its different class using the same techniques as in KNN, $x_{i,nh}$ is called "near-hit", $x_{i,nm}$ is called "near-miss". Next, update the weight of a feature A in W as described in Algorithm 1 [28, 29]. Repeat the random sampling steps for m times and get the average value of $W[A]$, $W[A]$ is the weight of feature A.

RELIEF Algorithm
Require: for each training instance, a vector of feature values and the class value
n ← number of training instances
a ← number of features
Parameter: m ← number of random training instances out of n used to update W
Initialize all feature weights $W[A] := 0.0$
For: $i := 1$ to m do
Randomly select a target instance R_i
find a nearest hit H and nearest miss M (instances)
For: $A := 1$ to a do
$W[A] := W[A] - diff(A, R_i, H)/m + diff(A, R_i, M)/m$
End For
End For
Return the weight vector W of feature scores that compute the quality of features

Algorithm 1: Pseudocode of the Relief algorithm

In Algorithm 1, $diff(x_a^j, x_b^j)$ depends on the type of feature j. For discrete feature j:

$$diff(x_a^j, x_b^j) = \begin{cases} 0, x_a^j = x_b^j \\ 1, otherwise, \end{cases}$$

for continuous feature j:

$$diff(x_a^j, x_b^j) = |x_a^j - x_b^j|.$$

Repeatedly operate for n times, then average the weights of each feature. Finally, choose the top k features for classification.

2.2 Machine Learning Classifiers

Machine learning classification algorithms are used to distinguish heart disease patients from healthy people. Five popular classifiers and their theoretical backgrounds are discussed briefly in this paper.

2.2.1 Random Forest

RF is a machine learning algorithm based on the ensemble of decision trees [30]. In traditional decision tree methods such as C4.5 and C5.0, all the features are used for generating the decision tree. In contrast, RF builds multiple decision trees and chooses the random subspaces of the features for each of them. Then, the votes of trees are aggregated and the class with the most votes is the prediction result [31].

2.2.2 Logistic Regression

LR is a generalized linear regression model [32]. Therefore, it is similar with multiple linear regression in many aspects. Usually, LR is used for binary classification problems where the predictive variable $y \in [0, 1]$, 0 is negative class and 1 is positive class. But it can also be used for multi-classification.

In order to distinguish heart disease patients from healthy people, a hypothesis $h(\theta) = \theta^T X$ is proposed. The threshold of classifier output is $h_\theta(x) = 0.5$, which is to say, if the value of hypothesis $h_\theta(x) \geq 0.5$, it will predict $y = 1$ which means that the person is a heart disease patient, otherwise the person is healthy. Hence, the prediction is done.

The sigmoid function of LR can be written as:

$$h_\theta(x) = \frac{1}{1 + e^{-z}},$$

where $z = \theta^T X$.

The cost function of LR can be written as:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} cost(y_i, y_i'),$$

where m is the number of instances to be predicted, y_i is the real class label of the ith instance, and y_i' is the predicted class label of the ith instance.

$$cost(y_i, y_i') = \begin{cases} 0, y_i = y_i' \\ 1, otherwise. \end{cases}$$

2.2.3 Support Vector Machine

Invented by Cortes and Vapnik [33], SVM is a supervised machine learning algorithm which has been widely used for classification problems [26, 34, 35]. The output of SVM is in the form of two classes in a binary classification problem, making it a non-probabilistic binary classifier [36]. SVM tries to find a linear maximum margin hyperplane that separates the instances.

Assume the hyperplane is $w^T x + b = 0$, where w is a dimensional coefficient vector, which is normal to the hyperplane of the surface, b is offset value from the origin, and x is dataset values. Obviously, the hyperplane is determined by w and b. The data points nearest to the hyperplane are called support vectors. In the linear case, w can be solved by introducing Lagrangian multiplier α_i. The solution of w can be written as:

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i,$$

where m is the number of support vectors and y_i are target labels to x. The linear discriminant function can be written as:

$$g(x) = sgn(\sum_{i=1}^{m} \alpha_i y_i x_i^T x + b),$$

sgn is the sign function that calculates the sign of a number, $sgn(x) = -1$ if $x < 0$, $sgn(x) = 0$ if $x = 0$, $sgn(x) = 1$ if $x > 0$. The nonlinear separation of data set is performed by using a kernel function. The discriminant function can be written as:

$$g(x) = sgn(\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b),$$

where $K(x_i, x)$ is the kernel function.

2.2.4 Extreme Learning Machine

ELM was first proposed by Huang et al. [37]. Similar to a single layer feed-forward neural network(SLFNN), ELM is also a simple neural network with a single hidden layer. However, unlike a traditional SLFNN, the hidden layer weights and bias of ELM are randomized and need not to tune, and the output layer weights of ELM are analytically determined through simple generalized inverse operations [37, 38].

2.2.5 K-Nearest Neighbor

KNN a supervised classification algorithm. Its procedure is as follows: when a new case is given, first search the database to find the k historical cases which are closest to the new case, namely k-nearest neighbors, and then these neighbors vote on the class label of the new case. If a class has the most nearest neighbors, the new case is determined to belong to the class [39]. The following formula is used to calculate the distance between two cases [40]:

$$d(x_i, x_j) = \sum_{q \in Q} w_q (x_{iq} - x_{jq})^2 + \sum_{c \in C} w_c L_c(x_{ic}, x_{jc}),$$

where Q is the set of quantitative features and C is the set of categorical features, L_c is an $M \times M$ symmetric matrix, w_q is the weight of feature q and w_c is the weight of feature c.

3 Proposed Framework

The proposed classification system consists of four main components: (1) preprocessing of data, (2) feature selection using Relief algorithm, (3) training of individual classifiers, and (4) prediction result generation of the ensemble classifier. A flow chart of the proposed system is shown in Figure 1. The main components of the system are described in the following subsections.

3.1 Data Preprocessing

The aim of data preprocessing is to obtain data from different heart disease data repositories and then process them in the appropriate format for the subsequent analysis [41]. The preprocessing phase involves missing-value imputation and data normalization.

3.1.1 Missing-value Imputation

Missing data in medical data sets must be handled carefully because they have a serious effect on the experimental results. Usually, researchers choose to replace the missing values with the mean/mode of the attribute depending on its type [23]. Mokeddem [41] used weighted KNN to calculate the missing values. In present study, features with missing values more than 50% of all instances are removed, then group mean instead of simple mean are used to substitute remaining missing values, as Bashir et al did in their study [35]. For example, if the case with a missing value is a patient, the mean value for patients is calculated and inserted in place of the missing value. In this way the class label is taken into consideration, thus the information offered by the dataset could be fully utilized.

3.1.2 Data Normalization

Before feature selection, the continuous features are normalized to ensure that they have the mean 0 and variance 1, thus the effects of different quantitative units are eliminated.

3.2 Feature Selection and Training of Individual Classifiers

In this phase, the dataset is randomly split into training set, validation set and test set. That is, 80% of the dataset is used for training, 10% is used for validation and 10% is used for testing purpose. The features are selected by the Relief algorithm on training set and the obtained result is a feature rank. A higher ranking means that the feature has stronger distinguishing quality and a higher weight [42]. In present study, different numbers of features are used to train individual classifiers on training set, and these generated models are tested on validation set to decide the best number of features, the rejected features are not used for subsequent modules and analysis.

3.3 Prediction Result Generation

The classification accuracy and misclassification cost (MC) of each classifier are taken into account during the process of generating the final prediction result. In present study, in order to compare the misclassification costs for the different classifiers conveniently, the value of the correct classification cost is set as 0, and the MC is split into two scenarios. In the first scenario, healthy people are diagnosed with heart disease, resulting in unnecessary and costly treatment. In the second scenario, heart disease patients are told that they are healthy, as a result they may miss the best time for treatment, which may cause the disease to deteriorate or even death. The cost matrix is presented in Table 2. Considering the different costs people have to pay for misclassification, we set $cost_1 = 5$ and $cost_2 = 1$. Afterwards, an index E is constructed to evaluate the performance of each classifier:

$$E_i = \frac{Accuracy_i + 1 - \frac{MC_i}{cost_1 + cost_2}}{2},$$

where $Accuracy_i$ represents the accuracy and MC_i represents the MC of ith classifier during the training phase (the formula to calculate the MC is presented in Section 4.2). E_i stands for the efficiency of ith classifier to improve the accuracy

and reduce the MC simultaneously. The weights of individual classifiers are based on E_i and they are calculated as:

$$w_i = \frac{E_i}{\sum_{i=1}^{n} E_i},$$

where n is the number of classifiers. Finally, the instances of the test set are imported into each classifier, and the outputs of ensemble classifier are the labels with the highest weighted vote[43].

4 Methods

In this paper, five individual classifiers including SVM, ELM, KNN, LR and RF are used for diagnosis of heart disease. Relief feature selection algorithm is used to select the most important features that have great influence on target predicted value. In order to evaluate the performance of ensemble model, various performance evaluation metrics such as misclassification cost (MC), G-mean, precision, specificity, recall and AUC (Area Under Curve) are used. In addition, data preprocessing techniques are applied to the heart disease datasets. In present study, the number of decision trees to build the RF is 50, the Gaussian kernel function is used in SVM, and the number of k is 5 in KNN.

In this research, important features selected by Relief algorithm are reported at first. Then, the performance of individual classifiers and the ensemble classifier are showed. Finally, a comparison is made between the performance of the proposed ensemble model and those of previous studies. The experiment is implemented on MATLAB 2018a platform, and the performance parameters of the executing host were Win 10, Inter (R) 1.80 GHz Core (TM) i5-8250U, X64, and 16 GB (RAM).

4.1 Datasets Description

Three different datasets are used in the proposed research, they are Statlog, Cleveland and Hungarian heart disease datasets from UCI machine learning repository. Statlog dataset consists of 270 instances, Cleveland dataset consists of 303 instances and Hungarian dataset consists of 294 instances. The number of heart disease patients in each dataset is presented in Table 2. The three datasets share the same feature set. Details of feature information are presented in Table 3.

4.2 Performance Evaluation Metrics

Various performance metrics are used to evaluate the performance of the classifiers in this study. In the confusion matrix, the classification result of a two-class problem is divided into four parts: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). Based on these error measures, E, MC, G-mean, precision, specificity, recall and AUC are used to evaluate the performance of different classifiers. As most health data sets have a non-balanced distribution of classes, accuracy is not a relevant metric in such cases, thus it is not used as an evaluation metric. The metrics are calculated as follows:

$$MC = \frac{FP \times cost_2 + FN \times cost_1}{TP + TN + FP + FN} \times 100\%, \quad (1)$$

$$G-mean=\sqrt{\frac{TP}{TP+FN}\times\frac{TN}{TN+FP}}\times 100\%, \tag{2}$$

$$Precision=\frac{TP}{TP+FP}\times 100\%, \tag{3}$$

$$Precision=\frac{TP}{TP+FP}\times 100\%, \tag{4}$$

$$Specificity=\frac{TN}{TN+FP}\times 100\%, \tag{5}$$

$$Recall=\frac{TP}{TP+FN}\times 100\%. \tag{6}$$

Ten-fold cross validation is used to obtain the final results. The ensemble model runs on each test set and processes each instance individually. The evaluation metrics of the ten folds are averaged to verify the superiority of the proposed ensemble classifier. The t-test is done on all three datasets to examine if the new method was statistically better as compared to single methods and check if the contribution of the Relief algorithm is significant.

5 Experimental Results

This section involves the exhibition of experimental results on different heart disease datasets.

5.1 Feature ranking on different datasets

Table 4 shows feature ranking on the three heart disease datasets. For Hungarian dataset, Slope, Ca and Thal are deleted during the process of missing-value imputation because these features have missing values more than 50% of all instances. Therefore, only ten features are ranked. Figure 2 illustrates how many times a certain feature is chosen to enter the best feature subset in the whole experiment. As we can see, sex, Cp, Exang, Slope, Ca and Thal are the most important features on Statlog dataset and Cleveland dataset, while sex, Cp, Trestbps, Exang and Oldpeak are the most important features on Hungarian dataset.

5.2 Performance on Statlog dataset

Table 5 indicates the comparison of performance evaluation metrics for the proposed ensemble with individual classifiers on Statlog dataset. The ensemble classifier performs the best on all the evaluation metrics, followed by SVM, and KNN

performs the worst. The result of t-test comparing the proposed ensemble and individual classifiers is shown in Table 6. It can be seen that the performance of proposed ensemble is significantly superior to individual classifiers on most of the metrics.

In order to investigate the contribution of Relief algorithm, experiments are done on Statlog dataset with all the features to make a comparison. The result is shown in Table 7. Compared with Table 5, the model with all the features is worse than that with feature subset chosen by Relief algorithm. Table 8 gives the result of statistical test between the two models, from which we can reach the conclusion that the difference is significant. In addition, it can be seen from Figure 2 that only 6 features on average are chosen by Relief algorithm for prediction, which reduces the computation largely.

5.3 Performance on Cleveland dataset

Table 9 shows the classification result of each classifier with reduced feature subset. The ensemble classifier performs the best on all the evaluation metrics while KNN performs the worst. The result of t-test comparing the proposed ensemble and individual classifiers is shown in Table 10. The ensemble classifier is obviously better than other classifiers on different metrics except for specificity.

The performance of the proposed model without Relief algorithm on Cleveland dataset is listed in Table 11. It can be concluded that the model performs worse than that with reduced feature subset, which indicates that there are irrelevant and distractive features. Table 12 shows the t-test result between the two models. As we can see, the classifiers gained significantly better performance with reduced feature subset. Besides, Relief algorithm has cut down the number of features to 8 on average, simplifying the calculation.

5.4 Performance on Hungarian dataset

Table 13 indicates the experimental results on Hungarian dataset with feature subset chosen by Relief algorithm. The proposed ensemble classifier has achieved the best performance on all the evaluation metrics. The t-test between the ensemble and each classifier is listed in Table 14. The ensemble is significantly superior to other classifiers on most of the metrics except specificity. This is because the proposed ensemble is cost-sensitive, its main aim is to identify patients as many as possible, thus the misclassification of healthy people is tolerable to a certain extent.

The performance of each classifier with all the features on Hungarian dataset is given in Table 15. The ensemble classifier still achieves the best performance on all the evaluation metrics. However, as shown in Table 16, the proposed model with Relief algorithm is not significantly better than the model without it on most of the indexes, this is because three important features were deleted in data preprocessing process, which weakened the effect of Relief algorithm.

5.5 Comparison of the Results with Other Studies

Table 17, 18 and 19 showed the comparison of our model and previous methods. As class imbalance is widespread in medical datasets, accuracy is not a good evaluation metric. Here, we use recall and specificity to make the comparison. Recall is

used to measure the percentage of distinguishing patients correctly, while specificity is used to measure the percentage of distinguishing healthy people correctly.

The results state that our proposed method obtains superior and promising results in classifying heart disease patients. Taken sensitivity and specificity together, the proposed ensemble classifier has better performance than most previous studies. In addition, most researchers did not take different kinds of misclassification costs into consideration, and the limitation is conquered in present study. Thus, we believe that the proposed model can be beneficial in aiding physicians in making better decisions.

6 Conclusions and Future Works

In this study, a cost-sensitive ensemble model based on five different classifiers is presented to assist the diagnosis of heart disease. The Statlog heart disease dataset, Cleveland heart disease dataset and Hungarian heart disease dataset are selected to test the model. The performance of classifiers are presented using different parameters such as E, MC, G-mean, precision, recall, specificity and AUC.

The main contributions of the proposed research are as follows:

(1) The proposed ensemble model is a novel combination of heterogeneous classifiers which had outstanding performance in previous studies. The limitations of a certain classifier are remedied by other classifiers in this model, which improves its performance.

(2)We have used E to combine the results of individual classifiers. The proposed ensemble model not only focuses on high classification accuracy, but also concerns the costs patients have to pay for misclassification.

(3)Compared with five individual classifiers and previous studies, the proposed ensemble model has achieved excellent classification results. The ensemble classifier gained significantly better performance than individual classifiers on all three heart disease datasets.

Kononenko [44] applied various machine learning techniques and compared the performance on eight medical datasets using five different parameters: performance, transparency, explanation, reduction, and missing data handling. While individual classifiers have shortcomings on some of these aspects, the ensemble model is able to overcome their deficiencies. For example, RF can generate explicit rules for decision making, and the basic idea of KNN is "to solve new problems by identifying and reusing previous similar cases based on the heuristic principle that similar problems have a high likelihood of having similar solutions" [45], which is easily understood by physicians. On the other hand, LR, SVM and ELM are more like a "black box", and physicians are willing to accept a "black box" classifier only when it outperforms a very large margin all other classifiers, including the physicians themselves, but such situation is highly improbable [44]. In addition, KNN is a lazy evaluation method while the other four are eager evaluation methods. Eager algorithm generates frequent itemset rules from a given data set and predicts a class for test instance based on multicriteria approach from selected frequent itemset rules [23]. If no matching is found, default prediction (i.e., the most frequent class in data set) is performed, which may not be correct. In contrast, lazy algorithm uses a richer hypothesis space, it makes judgment according to a small proportion of the instances

in the database, thus overcomes the limitation of eager algorithms. However, lazy algorithm uses more time for prediction, as multicriteria matching is performed for each instance in data set [46], while eager algorithm is able to generate the prediction results at a very fast speed after the training phase. From the above discussion, it can be concluded that the selected classifiers complement each other very well. In any scenario where one classifier has some limitations, the other classifier overcome them. As a result, better performance is achieved.

Moreover, the present study takes MC into consideration and tries to reduce it. Most traditional algorithms focus only on the classification accuracy, ignoring the cost patients have to pay for misclassification. But the diagnostic mistakes are of higher importance in the medical field, and the price of a false negative instance is clearly much higher than that of a false positive one. Aiming at this problem, the present study has adopted a new method to combine the prediction results of heterogeneous classifiers and significantly reduced the MC.

Compared with the state-of-the-art methods[47–50], the proposed model has certain advantages: (1) It is easily understood by less experienced clinical physicians, which makes it easier to implement. (2) Considering different kinds of misclassification cost makes the proposed model closer to reality. (3) This paper did not take accuracy as an evaluation metric, so the ensemble model is more suitable to imbalanced datasets. However, there are also shortages and limitations: (1) The experiment did not take training time into consideration. The ensemble model needs longer training time than individual classifiers. (2) The proposed approach doesn't include state-of-the-art techniques such as deep neural network and genetic algorithm. In the future, the time complexity of the proposed model will be investigated and optimized, and new algorithms can be incorporated into the proposed ensemble classifier to improve its performance.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Funding

This study was not funded.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Qi and Zhang designed research, performed research, analyzed data, and wrote the paper.

Acknowledgements

The authors acknowledge the editor and anonymous reviewers for their supportive works and insightful comments.

Availability of data and materials

The data used in this study is available in UCI Machine Learning Repository.

Author details
[1]College of Management and Economics, Tianjin University, Nankai District, 300072 Tianjin, PR China. [2]School of Mathematical Science, Hebei Normal University, Yuhua District, 050024 Shijiazhuang, PR China.

References

1. Heart disease. http://health.allrefer.com/health/heart-disease-info.html/Accessed:17.04.06
2. World Heart Federation Report. http://www.world-heart-federation.org/Accessed:01.12.16
3. for Cardiovascular Diseases, N.C.: The epidemic of heart disease. Encyclopedia of China Publishing House (2019)
4. Lopez-Sendon, J.: The heart failure epidemic. Medicographia **33**(2), 363–369 (2011)
5. Amato, F., Lopez, A., Pena-Mendez, E.M., Vanhara, P., Hampl, A., Havel, J.: Artificial neural networks in medical diagnosis. Journal of Applied Biomedicine **11**(2), 47–58 (2013)
6. Xu, M., Shen, J.: Information sharing system for heart disease emergence treatment based on an information fusion model. Industrial Engineering Journal **12**(4), 61–66 (2009)
7. Giri, D., Acharya, U.R., Martis, R.J., Sree, S.V., Lim, T.C., Thajudin Ahamed, V.I., Suri, J.S.: Automated diagnosis of coronary artery disease affected patients using lda, pca, ica and discrete wavelet transform. Knowledge-Based Systems **37**(2), 274–282 (2013)
8. Safdar, S., Zafar, S., Zafar, N., Khan, N.F.: Machine learning based decision support systems (dss) for heart disease diagnosis: a review. Artificial Intelligence Review **2017**, 1–27 (2017)
9. U Rajendra, A., Oliver, F., Vinitha, S., Swapna, . G., Roshan Joy, M., Nahrizul Adib, K., Suri, J.S.: Linear and nonlinear analysis of normal and cad-affected heart rate signals. Computer Methods & Programs in Biomedicine **113**(1), 55–68 (2014)
10. Mejia, O.A.V., Antunes, M.J., Goncharov, M., Dallan, L.R.P., Veronese, E., Lapenna, G.A., Lisboa, L.A.F., Dallan, L.A.O., Brandao, C.M.A., Zubelli, J., Tarasoutchi, F., Pomerantzeff, P.M.A., Jatene, F.B.: Predictive performance of six mortality risk scores and the development of a novel model in a prospective cohort of patients undergoing valve surgery secondary to rheumatic fever. PLoS ONE **2018**, 1–14 (2018)
11. Lukacs Krogager, M., Skals, R.K., Appel, E.V.R., Schnurr, T.M., Engelbrechtsen, L., Have, C.T., Pedersen, O., Engstrom, T., Roden, D.M., Gislason, G., Poulsen, H.E., Kober, L., Stender, S., Hansen, T., Grarup, N., Andersson, C., Torp-Pedersen, C., Weeke, P.E.: Hypertension genetic risk score is associated with burden of coronary heart disease among patients referred for coronary angiography. PLos One **13**(12), 1–17 (2018)
12. Tomar, D., Agarwal, S.: Feature selection based least square twin support vector machine for diagnosis of heart disease. International Journal of Bio-Science and Bio-Technology **6**, 69–82 (2014)
13. Subbulakshmi, C.V., Deepa, S.N.: Medical dataset classification: A machine learning paradigm integrating particle swarm optimization with extreme learning machine classifier. The scientific world journal **2015**, 1–12 (2015)
14. Jabbar, M.A., Deekshatulu, Chandra, P.: Heart disease classification using nearest neighbor classifier with feature subset selection. Computer Science & Telecommunications **2**, 47–54 (2013)
15. Dogan, M.V., Grumbach, I.M., Michaelson, J.J., Philibert, R.A.: Integrated genetic and epigenetic prediction of coronary heart disease in the framingham heart study. Plos One **13**(1), 1–18 (2018)
16. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.J., Sandhu, S., Guppy, K.H., Lee, S., Froelicher, V.: International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology **64**(5), 304–310 (1989)
17. Gokulnath, C.B., Shantharajah, S.P.: An optimized feature selection based on genetic approach and support vector machine for heart disease. Cluster Computing (4), 1–11 (2018)
18. Subbulakshmi, C.V., Deepa, S.N., Malathi, N.: Extreme learning machine for two category data classification. In: IEEE International Conference on Advanced Communication Control & Computing Technologies (2012)
19. Duch, W., Adamczak, R., K., G.: A new methodology of extraction, optimization and application of crisp and fuzzy logical rules. IEEE Transactions on Neural Networks **12**(2), 277–306 (2001)
20. Yingtsang, L.O., Fujita, H., Pai, T.: Prediction of coronary artery disease based on ensemble learning approaches and co-expressed observations. Journal of Mechanics in Medicine & Biology **16**(01), 1–10 (2016)
21. Eom, J.H., Kim, S.C., Zhang, B.T.: Aptacdss-e: A classifier ensemble-based clinical decision support system for cardiovascular disease level prediction. Expert Systems with Applications **34**(4), 2465–2479 (2008)
22. Das, R., Turkoglu, I., Sengur, A.: Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications **36**(4), 7675–7680 (2009)
23. Bashir, S., Qamar, U., Khan, F.H.: A multicriteria weighted vote-based classifier ensemble for heart disease prediction. Computational Intelligence **32**(4), 615–645 (2016)
24. Khened, M., Kollerathu, V.A., Krishnamurthi, G.: Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. Medical Image Analysis **51**, 21–45 (2018)
25. Krawczyk, B., Schaefer, G., Wozniak, M.: A hybrid cost-sensitive ensemble for imbalanced breast thermogram classification. Artificial Intelligence in Medicine **65**(3), 219–227 (2015)
26. Liu, N., Shen, J., Xu, M., Gan, D., Qi, E.S.: Improved cost-sensitive support vector machine classifier for breast cancer diagnosis. Mathematical Problems in Engineering **4**, 1–13 (2018)
27. Wei, Z., Junjie, C.: Relief feature selection and parameter optimization for support vector machine based on mixed kernel function. International Journal of Performability Engineering **14**(2), 280–289 (2018)
28. Ul Haq, A., Jian Ping, L., Memon, M.H., Nazir, S., Sun, R.: A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems **2018**, 1–21 (2018)
29. Urbanowicz, R.J., Meeker, M., Lacava, W., Olson, R.S., Moore, J.H.: Relief-based feature selection: Introduction and review. Journal of Biomedical Informatics **85**, 189–203 (2018)

30. Breiman, L.: Random forest. Machine Learning **45**, 5–32 (2001)
31. Hajialian, H., Toma, C.: Network anomaly detection by means of machine learning: Random forest approach with apache spark. Informatica Economica **22**(4), 89–98 (2018)
32. Larsen, K., Petersen, J.H., Budtz-Jorgensen, E., Endahl, L.: Interpreting parameters in the logistic regression model with random effects. Biometrics **56**(3), 909–914 (2015)
33. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (1995)
34. Davari, D.A., Khadem, S.E., Asl, B.M.: Automated diagnosis of coronary artery disease (cad) patients using optimized svm. Computer Methods & Programs in Biomedicine **138**, 117–126 (2017)
35. Bashir, S., Qamar, U., Khan, F.H.: Bagmoov: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting. Australas Phys Eng Sci Med **38**(2), 305–323 (2015)
36. Ghumbre, S., Patil, C., Ghatol, A.: Heart disease diagnosis using support vector machine. In: International Conference on Computer Science and Information Technology (ICCSIT), Pattaya, Thailand (2011)
37. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: Theory and applications. Neurocomputing **70**(1), 489–501 (2006)
38. Huang, G.B., Wang, D.H., Lan, Y.: Extreme learning machines: a survey. International Journal of Machine Learning & Cybernetics **2**(2), 107–122 (2011)
39. Wang, X., Li, H., Zhang, Q., Wang, R.: Predicting subcellular localization of apoptosis proteins combining go features of homologous proteins and distance weighted knn classifier. BioMed Research International **2016**(2), 1–8 (2016)
40. Uguroglu, S., Carbonell, J., Doyle, M., Biederman, R.: Cost-sensitive risk stratification in the diagnosis of heart disease. In: Twenty-sixth Aaai Conference on Artificial Intelligence (2012)
41. Mokeddem, S.A.: A fuzzy classification model for myocardial infarction risk assessment. Applied Intelligence (12), 1–18 (2017)
42. Zhang, L.X., Wang, J.X., Zhao, Y.N., Yang, Z.H.: A novel hybrid feature selection algorithm: using relieff estimation for ga-wrapper search. In: International Conference on Machine Learning & Cybernetics (2004)
43. Saha, S., Ekbal, A.: Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition. Data & Knowledge Engineering **85**(8), 15–39 (2013)
44. Kononenko, I.: Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine **23**(1), 89–109 (2001)
45. Ahmed, M.U., Begum, S., Olsson, E., Ning, X., Funk, P.: Case-based reasoning for medical and industrial decision support systems. Springer (2010)
46. Houeland, T.G., Aamodt, A.: An efficient hybrid classification algorithm - an example from palliative care. Springer **6679**, 197–204 (2011)
47. Ali, L., Khan, S.U., Golilarz, N.A., Yakubu, I., Nour, R.: A feature-driven decision support system for heart failure prediction based on χ^2 statistical model and gaussian naive bayes. Computational and Mathematical Methods in Medicine **2019**(4), 1–8 (2019)
48. Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., Khan, J.A.: An automated diagnostic system for heart disease prediction based on χ^2 statistical model and optimally configured deep neural network. IEEE Access, 1–1 (2019)
49. Ali, L., Niamat, A., Khan, J.A., Golilarz, N.A., Bukhari, S.A.C.: An optimized stacked support vector machines based expert system for the effective prediction of heart failure. IEEE Access **7**, 54007–54014 (2019)
50. SYED ARSLAN ALI, A.K.M.A.R.S.M.F.H.A. BASIT RAZA, KUMAR, Y.J.: An optimally configured and improved deep belief network (oci-dbn) approach for heart disease prediction based on ruzzo-tompa and stacked genetic algorithm. Digital Object Identifier **8**, 65947–65958 (2020)
51. Marateb HR, G.S.: A noninvasive method for coronary artery diseases diagnosis using a clinically-interpretable fuzzy rule-based system. J Res Med Sci **20**(3), 214–223 (2015)
52. Ceylan, R., Koyuncu, H.: A new breakpoint in hybrid particle swarm-neural network architecture: Individual boundary adjustment. International Journal of Information Technology & Decision Making, 1–31 (2016)
53. Das, R., Turkoglu, I., Sengur, A.: Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications **36**(4), 7675–7680 (2009)
54. Xiao, L., Wang, X., Qiang, S., Mo, Z., Zhu, Y., Wang, Q., Qian, W.: A hybrid classification system for heart disease diagnosis based on the rfrs method. Computational & Mathematical Methods in Medicine **2017**, 1–11 (2017)
55. Kahramanli, H., Allahverdi, N.: Design of a hybrid system for the diabetes and heart diseases. Expert Systems with Applications **35**(1-2), 82–89 (2008)
56. Shah, S.M.S., Batool, S., Khan, I., Ashraf, M.U., Abbas, S.H., Hussain, S.A.: Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis. Physica A Statistical Mechanics & Its Applications, 796–807 (2017)
57. Gorzaczany, M.B., Rudzinski, F.: Interpretable and accurate medical data classification - a multi-objective genetic-fuzzy optimization approach. Expert Systems with Applications **71** (2016)
58. Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., Moosaei, H., Yarifard, A.A.: Computer aided decision making for heart disease detection using hybrid neural network-genetic algorithm. Computer Methods & Programs in Biomedicine **141**(Complete), 19–26 (2017)
59. Leema, N., Nehemiah, H.K., Kannan, A.: Neural network classifier optimization using differential evolution with global information and back propagation algorithm for clinical datasets. Applied Soft Computing **49**, 834–844 (2016)
60. Mokeddem, S., Atmani, B.: Assessment of clinical decision support systems for predicting coronary heart disease. In: Fuzzy Systems: Concepts, Methodologies, Tools, and Applications, pp. 184–201. IGI Global, ??? (2016)

Figures

[scale=1]flow

Figure 1: Flowchart of the proposed ensemble classifier

[height=8cm,width=14cm]feature

Figure 2: Times picked for each feature on the three datasets

Tables

Table 1: The cost matrix used by the classifiers

Predicted	Reality	
	sick	healthy
sick	0	$cost_2$
healthy	$cost_1$	0

Table 2: Number of patients in each dataset

dataset	sick	healthy
Statlog	120	150
Cleveland	139	164
Hungarian	106	188

Table 3: Features of heart disease datasets

Feature	Description	Value
Age	Age in years	Continuous value
Sex	Gender	1 : male;0 : female
Cp	Chest Pain Type	1: typical angina 2: atypical angina 3: non-anginal pain 4: asymptomatic
Trestbps	Resting Blood Sugar	Continuous value in mm hg
Chol	Serum Cholestoral	Continuous value in mm/dl
Fbs	Fasting Blood Sugar	$0 :< 120$mg/dl $1 :> 120$mg/dl
Restecg	Resting ECG Results	0 : normal 1 : having ST-T wave abnormality 2 : probable or definite left ventricular hypertrophy
Thalach	Maximum heart rate achieved	Continuous value
Exang	Exercise induced angina	0 :no 1 :yes
Oldpeak	ST depression induced by exercise relative to rest	Continuous value
Slope	Slope of the peak exercise ST segment	1 =upsloping 2 =flat 3 =downsloping
Ca	Number of major vessels colored by flourosopy	0,1,2,3
Thal	Heart beat	3 :normal 6 :fixed defect 7 :reversable defect
Num	Predicted Class	$0, 1$

Table 4: Feature ranking on different datasets

Feature	Statlog	Cleveland	Hungarian
Age	9	9	7
Sex	4	4	2
Cp	1	1	1
Trestbps	8	8	5
Chol	13	13	6
Fbs	11	12	10
Restecg	7	7	8
Thalach	12	10	9
Exang	6	5	4
Oldpeak	10	11	3
Slope	5	6	\[1]
Ca	2	2	\
Thal	3	3	\

[1] \ means that feature doesn't exist in the dataset

Table 5: Experimental results on Statlog dataset with the best feature subset

Mean ± SD	RF	LR	SVM	ELM	KNN	Proposed ensemble
E(%)	87.53±5.39	87.87±6.82	88.67±5.02	82.81±5.54	76.94±11.33	**94.44±3.78**[*]
Precision(%)	83.70±6.58	84.07±8.01	84.81±6.40	78.15±6.64	70±15.37	**92.59±4.62**
Recall(%)	80.64±11.80	82.08±13.07	83.85±10.98	70.65±13.77	62.85±17.51	**92.15±7.10**
G-mean	83.14±7.54	83.79±8.19	84.41±7.10	76.65±8.05	68.40±15.63	**92.56±4.79**
MC(%)	51.85 ±26.07	50±34.67	44.81±23.78	75.19±29.63	96.67±44.56	**22.22±19.36**
Specificity(%)	86.13±6.17	86±6.58	85.45±7.83	84.29±8.43	75.18±16.55	**93.21±5.43**
AUC(%)	83.75±8.26	83.92±9.44	85.07±7.72	80.17±6.96	68.42±13.73	**92.08±5.51**

[*] The best result is bolded.

Table 6: t-test: Proposed Ensemble versus Individual Classifiers on Statlog dataset

		RF	LR	SVM	ELM	KNN
E	h[1]	1	1	1	1	1
	p[2]	0.0043	0.0184	0.01	5.14×10^{-5}	7.27×10^{-4}
Precision	h	1	1	1	1	1
	p	0.0030	0.0111	0.0065	3.57×10^{-5}	0.0011
Recall	h	1	0	0	1	1
	p	0.0218	0.0559	0.0716	7.34×10^{-4}	3.69×10^{-4}
G-mean	h	1	1	1	1	1
	p	0.0044	0.0108	0.0085	8.47×10^{-5}	7.37×10^{-4}
MC	h	1	1	1	1	1
	p	0.0105	0.0439	0.0322	2.45×10^{-4}	3.76×10^{-4}
Specificity	h	1	1	1	1	1
	p	0.0141	0.0159	0.0203	0.0128	0.0075
AUC	h	1	1	1	1	1
	p	0.0167	0.0203	0.0017	0.0053	2.76×10^{-4}

[1] the result of the test, $h = 1$ indicates that the null hypothesis can be rejected at the 5% level

[2] the probability of observing the given result by chance if the null hypothesis is true

Table 7: Experimental results on Statlog dataset with 13 features

Mean ± SD	RF	LR	SVM	ELM	KNN	Proposed ensemble
E(%)	71.76±7.44	77.16±4.53	68.49±6.06	77.31±8.11	66.70±4.26	**86.36±5.51**[*]
Precision(%)	65.19±14.96	73.70±7.77	68.15±10.03	61.48±29.28	59.26±11.05	**78.52±7.37**
Recall(%)	86.54±10.48	83.13±8.57	75.62±6.28	82.45±18.42	73.57±13.65	**92.56±8.19**
G-mean	82.18±9.64	83.72±14.18	76.29±7.45	82.60±14.51	76.35±18.16	**90.17±8.08**
MC(%)	75.12 ±9.10	56.30±7.77	62.69±25.27	41.12±33.75	85.19±43.82	**34.81±24.58**
Specificity(%)	78.05±7.26	84.32±8.97	76.96±16.40	82.81±8.72	79.23±17.11	**87.84±5.73**
AUC(%)	79.35±11.28	83.16±9.78	83.16±9.82	81.27±12.51	78.53±6.94	**87.99±8.39**

[*] The best result is bolded.

Table 8: t-test: Classifiers with feature subset versus Classifiers with 13 features on Statlog dataset

		RF	LR	SVM	ELM	KNN	Ensemble
E	h	1	1	1	1	1	1
	p	9.47×10^{-5}	0.0183	0.0319	4.56×10^{-4}	0.0341	0.0035
Precision	h	1	1	1	1	1	1
	p	0.0143	0.0433	0.0103	0.0011	0.0043	0.0296
Recall	h	0	1	1	1	0	0
	p	0.24	2.45×10^{-5}	0.0386	3.55×10^{-5}	0.2548	0.1826
G-mean	h	1	1	1	1	1	1
	p	0.0144	0.0058	0.0074	3.02×10^{-5}	4.63×10^{-4}	0.0017
MC	h	1	0	1	1	0	1
	p	0.0294	0.0791	0.0036	1.37×10^{-4}	0.0528	0.0059
Specificity	h	1	1	0	1	1	1
	p	0.0013	0.0156	0.1310	0.0282	0.0192	5.01×10^{-4}
AUC	h	1	1	1	1	1	1
	p	0.0151	0.0086	0.0129	0.0092	1.57×10^{-4}	5.28×10^{-4}

Table 9: Experimental results on Cleveland dataset with the best feature subset

Mean ± SD	RF	LR	SVM	ELM	KNN	Proposed ensemble
E(%)	86.78±6.15	86.53±6.75	86.50±5.89	84.19±7.59	79.44±9.05	**93.83±4.93***
Precision(%)	82.67±7.28	83.00±7.45	82.00±6.25	79.00±8.32	72.00±11.88	**88.67±5.49**
Recall(%)	80.26±14.28	78.02±16.41	81.20±15.12	77.86±19.94	73.70±14.34	**89.68±8.78**
G-mean	82.24±8.84	82.24±9.12	81.51±8.03	78.77±11.80	72.01±11.84	**90.77±6.71**
MC(%)	54.67 ±33.45	59.67±38.12	54.00±35.38	63.67±42.95	78.67±39.79	**22.00±15.61**
Specificity(%)	84.63±7.49	87.49±6.38	82.47±6.54	80.42±7.43	71.26±14.11	**89.31±5.13**
AUC(%)	81.53±8.75	81.99±9.38	80.91±8.14	79.99±11.05	70.53±12.65	**89.54±5.54**

* The best result is bolded.

Table 10: t-test: Proposed Ensemble versus Individual Classifiers on Cleveland dataset

		RF	LR	SVM	ELM	KNN
E	h	1	1	1	1	1
	p	6.42×10^{-4}	0.0024	2.00×10^{-4}	0.0014	3.41×10^{-4}
Precision	h	1	1	1	1	1
	p	0.0013	0.0028	4.02×10^{-4}	8.92×10^{-4}	1.60×10^{-4}
Recall	h	1	1	1	1	1
	p	6.83×10^{-4}	0.0033	0.0013	0.0065	0.0046
G-mean	h	1	1	1	1	1
	p	7.61×10^{-4}	0.0027	2.79×10^{-4}	0.0013	1.79×10^{-4}
MC	h	1	1	1	1	1
	p	8.66×10^{-4}	0.0037	7.18×10^{-4}	0.0046	0.0021
Specificity	h	0	0	0	1	1
	p	0.2108	0.6233	0.1169	0.0260	5.35×10^{-4}
AUC	h	1	1	1	1	1
	p	0.0021	0.0180	0.0019	0.0066	0.0024

Table 11: Experimental results on Cleveland dataset with 13 features

Mean ± SD	RF	LR	SVM	ELM	KNN	Proposed ensemble
E(%)	76.01±5.39	77.29±5.52	75.74±6.15	68.29±8.95	58.43±4.32	**82.07±6.00***
Precision(%)	74.23±6.41	76.84±5.14	75.16±7.47	65.54±11.57	50.26±6.74	**83.79±7.59**
Recall(%)	68.08±7.92	69.40±13.02	69.41±12.68	56.75±14.76	45.20±7.59	**75.88±11.08**
G-mean	71.05±6.75	73.59±6.58	71.61±7.07	61.45±12.32	49.71±6.20	**79.76±7.76**
MC(%)	87.19 ±21.18	81.61±27.83	82.08±29.68	114.60±37.19	152.39±19.74	**62.96±26.52**
Specificity(%)	74.50±9.02	79.31±9.11	74.80±8.20	67.32±11.32	49.20±11.80	**84.16±6.70**
AUC(%)	70.22±7.74	72.18±5.69	71.18±7.73	66.75±11.40	45.32±8.33	**79.53±8.24**

* The best result is bolded.

Table 12: t-test: Classifiers with feature subset versus Classifiers with 13 features on Cleveland dataset

		RF	LR	SVM	ELM	KNN	Ensemble
E	h	1	1	1	1	1	1
	p	1.23×10^{-4}	0.0014	2.85×10^{-4}	4.56×10^{-4}	8.17×10^{-6}	8.90×10^{-5}
Precision	h	1	1	1	1	1	1
	p	0.0059	0.0265	0.0307	0.0083	5.28×10^{-5}	0.0157
Recall	h	1	0	1	1	1	1
	p	0.0029	0.1415	0.0246	0.0038	2.74×10^{-4}	6.94×10^{-4}
G-mean	h	1	1	1	1	1	1
	p	7.02×10^{-4}	0.0085	0.0023	0.0021	9.16×10^{-6}	0.0013
MC	h	1	0	1	1	1	1
	p	0.0027	0.1090	0.0239	0.0055	2.66×10^{-4}	7.79×10^{-4}
Specificity	h	1	1	0	1	1	0
	p	0.0157	0.0478	0.0811	0.0088	3.54×10^{-4}	0.1434
AUC	h	1	1	1	1	1	1
	p	0.0024	0.0051	0.0058	0.0100	2.82×10^{-4}	0.0046

Table 13: Experimental results on Hungarian dataset with the best feature subset

Mean ± SD	RF	LR	SVM	ELM	KNN	Proposed ensemble
E(%)	80.43±5.37	82.07±7.12	78.91±5.61	80.40±6.86	75.43±8.64	**89.47±3.06***
Precision(%)	75.52±5.96	77.93±8.48	74.48±6.54	75.86±7.09	66.55±14.99	**89.31±4.44**
Recall(%)	60.19±16.84	62.08±15.89	53.38±17.93	59.42±19.49	61.36±19.71	**82.39±5.73**
G-mean	71.04±8.34	73.72±10.15	67.55±9.21	70.97±10.16	59.97±24.07	**82.95±4.63**
MC(%)	87.93 ±34.95	82.76±37.63	100.00±36.09	90.34±44.33	94.14±30.89	**38.28±12.10**
Specificity(%)	86.34±9.83	88.99±7.79	89.10±11.61	88.13±9.92	70.92±25.22	**92.02±5.76**
AUC(%)	74.07±9.16	76.31±10.87	71.96±10.98	74.59±9.55	69.07±9.98	**88.38±5.36**

* The best result is bolded.

Table 14: t-test: Proposed Ensemble versus Individual Classifiers on Hungarian dataset

		RF	LR	SVM	ELM	KNN
E	h	1	1	1	1	1
	p	5.62×10^{-5}	0.0023	2.29×10^{-5}	5.05×10^{-4}	1.64×10^{-4}
Precision	h	1	1	1	1	1
	p	2.03×10^{-5}	0.0022	2.20×10^{-5}	1.31×10^{-4}	8.45×10^{-4}
Recall	h	1	1	1	1	1
	p	0.0023	0.0028	5.17×10^{-4}	0.0046	0.0083
G-mean	h	1	1	1	1	1
	p	6.34×10^{-5}	0.0015	2.68×10^{-5}	3.72×10^{-4}	0.0050
MC	h	1	1	1	1	1
	p	0.0013	0.0046	3.29×10^{-4}	0.0047	1.97×10^{-4}
Specificity	h	0	0	0	0	1
	p	0.0507	0.1193	0.2511	0.1264	0.0257
AUC	h	1	1	1	1	1
	p	7.33×10^{-4}	0.0076	9.38×10^{-4}	0.0013	1.01×10^{-4}

Table 15: Experimental results on Hungarian dataset with 10 features

Mean ± SD	RF	LR	SVM	ELM	KNN	Proposed ensemble
E(%)	72.73±6.29	73.85±7.06	72.72±6.78	69.94±8.26	60.09±10.59	**79.87±7.32***
Precision(%)	72.72±8.17	73.38±8.14	71.78±8.31	69.18±10.08	53.77±13.27	**80.89±7.89**
Recall(%)	49.00±16.03	52.92±14.85	44.30±17.06	44.39±20.61	37.77±18.40	**66.38±14.13**
G-mean	62.75±11.18	65.96±10.60	60.44±12.46	58.39±14.78	45.48±14.87	**75.75±9.22**
MC(%)	109.40 ±31.01	103.24±32.31	118.24±33.24	123.00±42.83	148.60±48.07	**74.08±32.11**
Specificity(%)	82.62±5.75	83.40±5.22	85.57±5.62	80.65±8.26	59.28±13.55	**87.31±3.60**
AUC(%)	67.38±10.99	68.59±10.98	65.43±10.99	61.67±13.98	50.81±15.55	**77.64±8.31**

* The best result is bolded.

Table 16: t-test: Classifiers with feature subset versus Classifiers with 10 features on Hungarian dataset

		RF	LR	SVM	ELM	KNN	Ensemble
E	h	1	1	1	1	1	1
	p	0.0088	0.0185	0.0398	0.0066	0.0024	5.85×10^{-4}
Precision	h	0	0	0	0	0	1
	p	0.3942	0.2363	0.4308	0.1077	0.0589	0.0106
Recall	h	0	0	0	0	1	1
	p	0.1458	0.1998	0.2613	0.1112	0.0127	0.0062
G-mean	h	0	0	0	1	0	1
	p	0.0780	0.1117	0.1653	0.0414	0.1261	0.0024
MC	h	0	0	0	0	1	1
	p	0.1636	0.2085	0.2551	0.1112	0.0085	0.0067
Specificity	h	0	0	0	0	0	1
	p	0.3176	0.0782	0.4021	0.0837	0.2472	0.0069
AUC	h	0	0	0	1	1	1
	p	0.1568	0.1317	0.2005	0.0283	0.0068	0.0036

Table 17: Comparison of the proposed system outcome with previous researches for Statlog dataset

Author	Method	Recall(%)	Specificity (%)
Present study	Ensemble classifier	92.15	93.21
Marateb and Goudarzi[51]	Naive Bayes	78.51	88.74
Bashir et al[35]	BagMOOV	73.47	91.01
Ceylan and Koyuncu[52]	PSO neural network	80.83	89.33
Mokeddem and Ahmed[41]	Fuzzy classification model	89.17	84.00
Das et al[53]	Neural network ensemble	80.95	95.91
Xiao et al[54]	Heuristic Rough Set	92.33	87.50
Bashir et al[23]	Ensemble model	87.50	87.27

Table 18: Comparison of the proposed system outcome with previous researches for Cleveland dataset

Author	Method	Recall(%)	Specificity (%)
Present study	Ensemble classifier	89.68	89.31
Kahramanli and Allahverdi[55]	Hybrid neural network	93	78.5
Shah et al[56]	PPCA*+ SVM	75	90.57
Marian and Filip [57]	Fuzzy rule-based classification	84.70	92.90
Ali et al[47]	Gaussian Naive Bayes classifier	87.80	97.95
Ali et al[48]	Deep neural network	85.36	100
Ali et al[49]	Hybrid SVM	82.92	100
Ali et al[50]	Deep belief network	96.03	93.15
Arabasadi et al[58]	Hybrid neural network-genetic algorithm	88	91
Mokeddem and Ahmed[41]	Fuzzy classification model	87.39	94.38
Bashir et al[23]	Ensemble model	73.68	92.86
Leema et al[59]	Differential Evolution + BPNN	82.35	92.31
Mokeddem and Atmani[60]	Decision Tree + Fuzzy Inference System	92.44	96.18

* Probabilistic Principal Component Analysis.

Table 19: Comparison of the proposed system outcome with previous researches for Hungarian dataset

Author	Method	Recall(%)	Specificity (%)
Present study	Ensemble classifier	82.39	92.02
Shah et al[56]	PPCA + SVM	80.43	88.42
Arabasadi et al[58]	Hybrid neural network-genetic algorithm	85	88
Mokeddem and Ahmed[41]	Fuzzy classification model	82.98	90.57
Mokeddem and Atmani[60]	Decision Tree + Fuzzy Inference System	90.42	79.24