Sādhanā Vol. 42, No. 7, July 2017, pp. 1135–1141
DOI 10.1007/s12046-017-0683-z

Prediction of cardiac arrest recurrence using ensemble classifiers

NACHIKET TAPAS*, TUSHAR LONE, DAMODAR REDDY and VENKATANARESH KUPPILI

Department of Computer Science and Engineering, National Institute of Technology Goa, Farmagudi, Ponda 403401, India
e-mail: nachikettapas@gmail.com

MS received 21 June 2016; revised 24 November 2016; accepted 17 January 2017

Abstract. Inability of a heart to contract effectually or its failure to contract prevents blood from circulating efficiently, causing circulatory arrest or cardiac arrest or cardiopulmonary arrest. The unexpected cardiac arrest is medically referred to as sudden cardiac arrest (SCA). Poor survival rate of patients with SCA is one of the most ubiquitous health care problems today. Recent studies show that heart-rate-derived features can act as early predictors of SCA. Addition of angiographic and electrophysiological features can increase the robustness of the prediction system. Early warning has the capability of saving many lives. Risk of recurrent terminal cardiac arrest is high for out-of-hospital survivors. Foregoing studies indicate that recurrent cardiac events are time dependent and, while in clinical follow-up, are highly probable, predominantly in early phase. In this paper, we observe the changing risk of and changing influence of various clinical, angiographic and electrophysiological parameters on subsequent cardiac arrest recurrence with time. Various medical and synthetic datasets such as ECG dataset from PhysioNet, Pima Indian Diabetes dataset from UCI Machine Learning Repository and gene expression dataset from GEO are used, which are unique as compared with related works. Various classifiers such as LogitBoost with simple regression function, random forest and multilayer perceptron are used for recurrence risk prediction. Collection of these classifiers together forms the ensemble classifiers. Classifiers are compared based on various measures like accuracy and precision. Based on the classification, risk scores are calculated using logistic regression with backward elimination. The proposed method is used for final risk estimation. The same datasets are used for risk score calculation model development. Experimental results are found to be encouraging.

Keywords. Heart rate variability; machine learning; ensemble classifier; sudden cardiac arrest.

1. Introduction

In general population, for every 100,000 people, 50–100 suffer from sudden cardiac death (SCD) [1] based on studies with multiple sources in the United States [2, 3], Netherlands [4], Ireland [5] and China [6]. In order to more accurately estimate the severity of SCD incidence, multiple sources of monitoring are required. Advances in genetics can also help us in forming highly focussed groups among general population. In India, based on verbal autopsy conducted by Rao *et al* [7] in 2012 with patient condition resulting in SCD, it was found that 10.3% of the overall mortality was due to SCD. In comparison with the studies conducted in United States and Europe, with 75 years as mean age, the SCD patients in India had a lower mean age of 60 years.

When a heart becomes non-functional due to stoppage of electrical activity, medically the condition is known as cardiac arrest [8]. The death caused due to cardiac arrest is known as cardiac death. The risk factors for SCD are similar to those of coronary artery disease and include (1) diabetes, (2) family history, (3) blood pressure, (4) smoking, (5) lack of physical exercise, etc. [9]. Using these risk factors, we can improve the accuracy of the prediction model.

In a combination of classifiers such that the results of individual classifiers are improved, such classifiers are known as ensemble classifiers. The goal of ensemble-based classification is to generate more precise, certain and accurate system results. Ensemble classifiers can be constructed using various techniques such as (1) single learning method with different subsets of data, (2) single learning method with varying parameters and (3) multiple learning methods. Bagging and boosting are two techniques of (1) type. In bagging, a training set of size p is constructed by selecting p samples with replacement from complete data. The process is repeated several times and each iteration results in a new classifier. The final classification is decided

*For correspondence

based on majority voting. In boosting, the aim is to train the classifier to identify those instances that were not classified properly. Thus, after each iteration, the training set is constructed to involve samples not properly classified in previous iteration.

Risk score is estimated using alternate weighing procedures with logistic regression. Those parameters with significance more than 20% participate in the estimation. The scoring systems are then scaled using different methods like dividing the coefficient by maximum value and rounding to the nearest integer. Other methods include multiplying the coefficient by 10 and then rounding to the nearest integer.

In this study, we propose an ensemble-classifier-based approach involving different risk factors, thereby making risk prediction accurate. The datasets involved are from various sources like ECG data from PhysioNet, gene expression data from GEO and diabetes and blood pressure data from UCI machine repository. For recurrence prediction, we generated synthetic dataset based on ECG data.

The organization of text is as follows. Section 2 contains related work. Section 3 consists of proposed system for prediction. Section 4 contains experimental results on various datasets. Section 5 is conclusion.

2. Related work

In this section, we review some existing work related to cardiac arrest recurrence prediction and risk score calculation.

2.1 *Time-dependent risk of and predictors for cardiac arrest recurrence*

This study is performed on 142 consecutive patients within 3 weeks of surviving out-of-hospital cardiac arrest. Electrophysiologic evaluation was conducted between October 1980 and June 1987. It was followed up to June 1988. Patients having coronary artery disease were selected for this study (101 patients). All patients required external direct-current counter shock for restoration of stable-cardiac rhythm.

In this study, feature vectors (table 1) [10] representing clinical characteristics are considered. Apart from standard features like age and sex, features include how many times patient has suffered from myocardial infraction, time gap between cardiac arrest and myocardial infraction, type of arrhythmia, ejection fraction, etc.; 81% had ventricular fibrillation rhythm while 19% had ventricular tachycardia rhythm at the time of cardiac arrest; 81% suffered from single episode of myocardial infarction while 21% had more than one episode. Cardiac catheterization was performed in 94%. Reduction greater than 70% is considered significant; coronary lesions were defined as greater than 70% reduction of the luminal diameter. One-vessel disease was detected in 29%, two-vessel disease in 37% and three-vessel disease in 34%. Segmental left ventricular wall motion abnormalities and ejection fraction were determined angiographically in 94% patients and by radionuclide-gated scan in 5%; 32% patients had a left ventricular aneurysm [11].

Table 1. Clinical characteristics of patients.

Characteristics
Age (year)
Sex (male/female)
Presented arrhythmia
Ventricular fibrillation
Ventricular tachycardia
Previous MI
More than one MI
Time duration from MI to CA
>4 months
<4 months
Congestive heart failure
Involved coronary vessels (n)
One vessel
Two vessels
Three vessels
Left ventricular aneurysm
Ejection fraction (%)
$\geq 35\%$
<35%

Feature vectors are compared using χ^2 test with Yates' correction or Student's unpaired t-test. Threshold for significance is set to predictive values (p values) less than 0.05. Cardiac arrest recurrence is predicted using actuarial curves and actuarial rates that are calculated using the Kaplan–Meier method [10]. The Mantel–Cox statistic was used to compare actuarial recurrence rates of the groups [12]. This analysis was used to identify significant variables with predictive value ($p < 0.05$). Actuarial recurrence rate was found to be 11.2% during first 6 months and decreased to less than 4% after 6 months.

Analysis was done on lower risk subsequent phase (>6 months) and high-risk early phase (6 months) using multivariate Cox proportional hazards analysis. An ejection fraction less than 35% was identified as a significant predictor for cardiac arrest recurrence during first 6 months [10]. Three variables: number of coronary vessels involved, congestive heart failure and persistent inducibility of VT, were found out to be significant predictors of cardiac arrest recurrence after 6 months [10].

2.2 *Risk score estimation of diabetic retinopathy*

Vaitheeswaran *et al* [16] used multinomial logistic regression with backward elimination for risk score calculation. The general form of logistic regression is as follows:

$$\eta_i = \text{logit}(odds_i) = \log\left(\frac{q_i}{1-q_i}\right) \quad (1)$$

where

$$q_i = X_i' * \beta + \epsilon_i\,. \quad (2)$$

The variable η_i is a Bernoulli random variable that is either 0 or 1; $X_i = \{1, X_{i1}, X_{i2}, \ldots, X_{ip}\}$ where each element is an attribute of feature vector; β is regression coefficient based on regression model and ϵ_i is error.

For point-based systems, different methods have different weights W_i associated, which are used for scores calculation using the logistic regression method. Description of existing methods and proposed method is shown in table 2. The standard method (SM) estimates risk scores without any round off for decimal places and thus final calculated score should be as close to SM as possible. Method 1 mentioned in [13] multiplies 10 to β coefficients and rounds to the nearest integer to get the weight for final risk calculation. Method 2 in [14] divides each β coefficient with minimum β, multiplying it by 2 and rounding it to the nearest integer. Method 3 in [15] takes mean of two smallest β and divides each β with the mean. The result is then multiplied with 2 and rounded to the nearest integer. In method 4 [16], each β coefficient is divided by β with the maximum value of Wald statistics. The result is multiplied by 100 and rounded to the nearest integer. In method 5 of [16] each β is divided by the maximum value of β and multiplied by 100 and rounded to the nearest integer.

3. Proposed solution

The proposed solution consists of data sources identified as risk factors for cardiac arrest. The datasets are preprocessed and are individually used to train logistic regression and random forest classifiers. The logistic regression with backward elimination is used to derive model for risk score.

Table 2. Existing and proposed method description for medical risk estimations.

Method	Source	W_i
M_1	Gulmer *et al* [13]	$10\beta_i$
M_2	Chen *et al* [14]	$\left[2\left(\frac{\beta_i}{\min\lvert\beta_i\rvert}\right)\right]$
M_3	Sugioka *et al* [15]	$\left[2\left(\frac{\beta_i}{\text{Mean}(\beta_p,\beta_q)}\right)\right]$
M_4	Vaitheeswaran *et al* [16]	$\left[100\left(\frac{\beta_i}{\beta_j}\right)\right]$
M_5	Vaitheeswaran *et al* [16]	$\left[100\left(\frac{\beta_i}{\max\lvert\beta_i\rvert}\right)\right]$

3.1 *Dataset*

We considered dataset from various sources like Pima Indian Diabetes dataset and cuff-less blood pressure estimation dataset from UCI machine repository and gene expression dataset from GEO. We also synthetically generated HRV dataset with data having multiple cardiac arrests using ECG data from PhysioNet.

Pre-processing is required for gene expression and ECG dataset. Gene expression datasets are preprocessed using Tophat and ECG dataset is preprocessed using Kubios HRV. In the gene expression dataset, after preprocessing, we get differentially expressed genes, which we use to train the classifier. For HRV, the processed ECG data are used to extract HRV parameters with class label. Synthetic dataset is generated by keeping the same distribution as that of the original data and creating others classes with diminishing HRV parameters. The data were generated using Matlab version R2014a.

3.2 *Classifiers*

The selection of classifier is done based on accuracy and complexity of the classifier. Since the amount of data is less, using a complex model will result in over-fitting. Based on this, simple models with high accuracy are selected. In order to improve accuracy, ensemble techniques (bagging, boosting) are used.

The logistic regression classifier uses the boosting technique while the random forest classifier is based on the bagging technique. The classification results are evaluated based on various measures like accuracy, precision, specificity, sensitivity, f-score and area under ROC curve. Also, various training–testing splits are used to derive the best split so that minimum number of training samples are required.

3.3 *Risk score*

A Wald test is used to test the statistical significance of each coefficient β in the model [17]. Wald statistics determines correlation between two quantities. Thus higher Wald statistics for a feature shows high correlation between dependent variable and that feature. The general approach is to use β coefficients from logistic regression to estimate weights or criticality of a particular feature. Higher the weight, more impact the change in that feature holds [13–16]. Vaitheeswaran *et al* [16] exploited the Wald statistics to find features with higher impact.

The proposed risk score model (M_6) uses weights such that each β of logistic regression is divided by mean of two βs that have highest Wald characteristics, i.e., mean of two β coefficients highly correlated to the dependent variable. Since β with the highest Wald statistics are chosen, the dependent variable is highly correlated to the two features. Thus results from the proposed method will be much more closer to actual predicted result.

Table 3. Wald statistics and β value for each attribute of diabetes dataset.

Attribute	Wald	β
A1	14.7467	0.1232
A2	89.8968	0.0352
A3	6.4537	−0.0133
A4	0.0080	0.0006
A5	1.7485	−0.0012
A6	35.3470	0.0897
A7	9.9829	0.9452
A8	2.5372	0.0149

$$W_i = \left[100\left(\frac{\beta_i}{\text{mean}(\beta_r, \beta_s)}\right)\right] \quad (3)$$

where β_r and β_s are β coefficients with the maximum value of Wald statistics. For example, table 3 shows Wald statistics and β of attributes of diabetes dataset; β_r is choosen to be that β of the attribute for which Wald statistics is the maximum. Similarly, β_s is choosen to be that β of the attribute for which Wald statistics is the second maximum. Thus, β_r is choosen to be 0.0352 corresponding to A2 with maximum Wald statistics. Similarly, β_s is choosen to be 0.0897 corresponding to A6 with the next maximum Wald statistics.

Another risk score model (M_7) uses weights such that each β of logistic regression is divided by square of mean of two βs that have the highest Wald characteristics, i.e., mean of two β coefficients highly correlated to the dependent variable.

$$W_i = \left[100\left(\frac{\beta_i}{(\text{mean}(\beta_r, \beta_s))^2}\right)\right]. \quad (4)$$

For comparing the proposed method with other methods, we used Pima Indian Diabetes dataset to generate the β coefficients. The dataset consists of 768 instances with 8 attributes. Regression was performed for class label as the dependent variable. We found that plasma glucose concentration and body mass index were the features with high Wald statistics score. Weights were calculated for individual systems based on the approach suggested. Comparison was done for some sample patients. Table 4 compares various scoring methods discussed earlier. SM does not contain any error due to rounding. We find that the results from SM are closest to scores predicted by the proposed method.

Table 4. Risk score for a particular patient using different methods.

Method	Score	Percent
SM	0.00100	1.26618
M_1	0.00000	0.00000
M_2	2.00000	1.26789
M_3	2.00000	1.26789
M_4	3.00000	1.25930
M_5	1.00000	1.55039
M_6 (proposed)	2.00000	1.26711
M_7 (proposed)	26.00000	1.26641

3.4 *Complexity analysis*

The proposed solution is presented in figure 1. The proposed method consists of two phases: training the classifier and risk score calculation. The worst case complexity for training a logistic regression classifier is $O(C^2N)$ where C is the number of features in the dataset and N is the number of training samples. The worst case complexity for training a random forest classifier is $O(PQN\log(N))$ where P is the number or trees in the forest, Q is the number of features to be sampled at each node and N is the number of training samples. For weight calculation in risk score formulation, we again use logistic regression and thus worst case complexity is $O(C^2N)$. Thus overall running time of the proposed system is almost linear.

4. Experimental results

The proposed solution was tested with different learning algorithms and different train–test splits for various measures. The measures include accuracy, precision, sensitivity, specificity, *f*-score and area under ROC curve.

4.1 *Experimental set-up*

All experiments are carried out on a Windows 7 operating system, Intel Core i7 4790 CPU with 8GB RAM. For

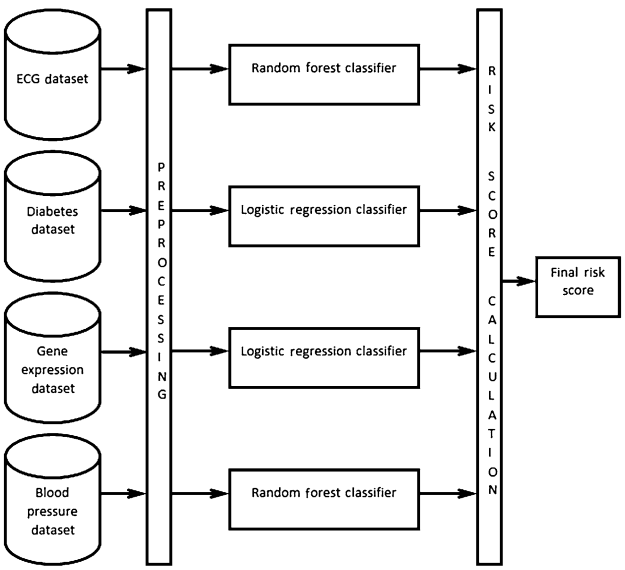

Figure 1. Proposed system.

implementation and experimentation, Eclipse Mars with Java, Weka 3.8 and Matlab version R2014a is used.

4.2 *Accuracy*

Table 5 compares the result of classification accuracy for blood pressure dataset using 70–30% train vs test split. Comparison is done between accuracy of various algorithms like MultiLayer Perceptron (MLP), Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR) and Radial Basis Function Neural Network (RBF). The training and testing cycle is repeated 10 times and average with standard deviation is presented in the table as well. The blood pressure dataset consists of 400 records. Thus 280 samples are used for training the classifiers while 120 samples are used as the test dataset.

Table 6 compares the result of classification accuracy for diabetes dataset using 70–30% train vs test split. The same set of machine learning algorithms is used as in the previous dataset. The diabetes dataset consists of 768 records. Thus 538 samples are used for training the classifiers while 230 samples are used as the test dataset.

Similarly tables 7 and 8 compare the result of classification accuracy for gene expression dataset and heart rate variability dataset, respectively, using 70–30% train vs test split. The gene expression dataset consists of 230 records. Thus 161 samples are used for training the classifiers while 69 samples are used as the test dataset. The heart rate variability dataset consists of 928 records. Thus 650 samples are used for training the classifiers while 278 samples are used as the test dataset.

Table 5. Accuracy statistics for 70–30 train–test split for blood pressure dataset.

Iteration	MLP	NB	RF	LR	SVM	RBF
1	0.92	0.97	0.99	0.98	0.34	0.94
2	0.93	0.93	1.00	1.00	0.28	0.91
3	0.93	0.98	0.99	0.99	0.38	0.82
4	0.93	0.99	0.98	0.98	0.31	0.95
5	0.95	1.00	0.98	1.00	0.33	0.94
6	0.96	1.00	0.99	1.00	0.38	0.92
7	0.97	0.97	0.98	0.98	0.28	0.96
8	0.98	0.97	0.99	0.99	0.37	0.95
9	0.98	1.00	1.00	1.00	0.33	0.94
10	1.00	0.97	1.00	1.00	0.32	0.85
Avg	0.96	0.98	0.99	0.99	0.33	0.92
Std	0.03	0.02	0.01	0.01	0.04	0.05

Table 6. Accuracy statistics for 70–30 train–test split for diabetes dataset.

Iteration	MLP	NB	RF	LR	SVM	RBF
1	0.70	0.70	0.73	0.72	0.65	0.71
2	0.71	0.76	0.76	0.75	0.64	0.75
3	0.71	0.77	0.77	0.79	0.69	0.77
4	0.73	0.80	0.78	0.80	0.65	0.79
5	0.73	0.76	0.80	0.78	0.66	0.78
6	0.77	0.76	0.77	0.79	0.67	0.76
7	0.77	0.78	0.80	0.79	0.65	0.79
8	0.77	0.77	0.76	0.74	0.63	0.73
9	0.78	0.77	0.77	0.78	0.63	0.77
10	0.80	0.83	0.83	0.80	0.67	0.81
Avg	0.75	0.77	0.78	0.77	0.65	0.77
Std	0.04	0.03	0.03	0.03	0.02	0.03

Table 7. Accuracy statistics for 70–30 train–test split for gene expression dataset.

Iteration	MLP	NB	RF	LR	SVM	RBF
1	0.93	0.93	0.94	0.93	0.61	0.93
2	0.94	0.93	0.94	0.94	0.54	0.94
3	0.96	0.93	0.96	0.90	0.49	0.93
4	0.97	0.94	0.94	0.93	0.58	0.94
5	0.97	0.96	0.97	0.97	0.52	0.97
6	0.97	0.96	0.99	0.97	0.61	0.96
7	0.97	0.97	0.99	0.97	0.57	0.99
8	0.99	0.94	0.97	0.97	0.61	0.99
9	0.99	0.97	0.97	0.97	0.59	0.99
10	1.00	0.99	0.99	0.99	0.61	1.00
Avg	0.97	0.95	0.97	0.95	0.57	0.96
Std	0.02	0.02	0.02	0.03	0.04	0.03

4.3 *Precision*

Precision can be defined as positive prediction power of the model, i.e., how well the model is able to classify positive or disease samples. Better precision implies high classification rate for identifying diseased samples. Tables 9–12 compare the results of precision for blood pressure, diabetes, gene expression and heart rate variability dataset, respectively, using 70–30% train vs test split. Comparison

Table 8. Accuracy statistics for 70–30 train–test split for heart rate variability dataset.

Iteration	MLP	NB	RF	LR	SVM	RBF
1	0.93	0.93	0.94	0.93	0.61	0.93
2	0.94	0.93	0.94	0.94	0.54	0.94
3	0.96	0.93	0.96	0.90	0.49	0.93
4	0.97	0.94	0.94	0.93	0.58	0.94
5	0.97	0.96	0.97	0.97	0.52	0.97
6	0.97	0.96	0.99	0.97	0.61	0.96
7	0.97	0.97	0.99	0.97	0.57	0.99
8	0.99	0.94	0.97	0.97	0.61	0.99
9	0.99	0.97	0.97	0.97	0.59	0.99
10	1.00	0.99	0.99	0.99	0.61	1.00
Avg	0.97	0.95	0.97	0.95	0.57	0.96
Std	0.02	0.02	0.02	0.03	0.04	0.03

Table 9. Precision statistics for 70–30 train–test split for blood pressure dataset.

Iteration	MLP	NB	RF	LR	SVM	RBF
1	0.89	0.96	0.99	0.98	0.09	0.94
2	0.94	0.99	0.98	0.98	0.08	0.94
3	0.94	1.00	0.99	1.00	0.08	0.92
4	0.94	1.00	0.99	1.00	0.09	0.94
5	0.95	0.96	0.98	0.98	0.10	0.65
6	0.95	0.94	1.00	1.00	0.07	0.92
7	0.97	1.00	1.00	1.00	0.08	0.95
8	0.98	0.96	0.99	0.99	0.07	0.94
9	0.99	0.96	0.99	0.99	0.09	0.93
10	1.00	0.98	1.00	1.00	0.08	0.64
Avg	0.95	0.97	0.99	0.99	0.08	0.88
Std	0.03	0.02	0.01	0.01	0.01	0.12

Table 10. Precision statistics for 70–30 train–test split for diabetes dataset.

Iteration	MLP	NB	RF	LR	SVM	RBF
1	0.68	0.73	0.73	0.77	0.34	0.74
2	0.68	0.67	0.71	0.71	0.32	0.68
3	0.69	0.74	0.74	0.73	0.32	0.73
4	0.70	0.78	0.76	0.78	0.33	0.77
5	0.71	0.74	0.78	0.77	0.33	0.76
6	0.73	0.73	0.73	0.77	0.34	0.72
7	0.75	0.76	0.78	0.77	0.33	0.77
8	0.76	0.76	0.75	0.73	0.32	0.72
9	0.79	0.82	0.82	0.80	0.33	0.80
10	0.79	0.76	0.76	0.79	0.32	0.76
Avg	0.73	0.75	0.76	0.76	0.33	0.74
Std	0.04	0.04	0.03	0.03	0.01	0.03

Table 11. Precision statistics for 70–30 train–test split for gene expression dataset.

Iteration	MLP	NB	RF	LR	SVM	RBF
1	0.92	0.92	0.94	0.92	0.30	0.92
2	0.94	0.93	0.94	0.94	0.27	0.94
3	0.96	0.93	0.96	0.90	0.25	0.93
4	0.97	0.96	0.99	0.98	0.30	0.96
5	0.97	0.94	0.94	0.92	0.29	0.94
6	0.97	0.96	0.97	0.97	0.26	0.97
7	0.98	0.97	0.99	0.97	0.28	0.99
8	0.99	0.98	0.98	0.98	0.30	0.99
9	0.99	0.94	0.97	0.97	0.30	0.99
10	1.00	0.98	0.99	0.99	0.30	1.00
Avg	0.97	0.95	0.97	0.95	0.29	0.96
Std	0.02	0.02	0.02	0.03	0.02	0.03

is done between precision of various algorithms like MLP, NB, RF, LR, SVM and RBF. The training and testing cycle is repeated 10 times and the average with standard deviation is presented in the table as well.

Table 12. Precision statistics for 70–30 train–test split for heart rate variability dataset.

Iteration	MLP	NB	RF	LR	SVM	RBF
1	0.82	0.91	0.86	0.82	0.02	0.79
2	0.82	0.89	0.89	0.86	0.02	0.37
3	0.83	0.90	0.89	0.86	0.02	0.93
4	0.85	0.90	0.91	0.85	0.01	0.79
5	0.85	0.88	0.87	0.83	0.02	0.90
6	0.85	0.89	0.89	0.82	0.01	0.93
7	0.86	0.90	0.90	0.82	0.01	0.79
8	0.86	0.87	0.89	0.83	0.02	0.57
9	0.87	0.88	0.87	0.82	0.02	0.52
10	0.88	0.90	0.90	0.83	0.01	0.40
Avg	0.85	0.89	0.89	0.83	0.01	0.70
Std	0.02	0.01	0.02	0.02	0.00	0.21

The configuration of datasets remains the same. Out of 400 records of blood pressure dataset, 280 samples are used for training and 120 for testing. Out of 768 records for diabetes dataset, 538 samples are used for training and 230 samples are used for testing; 230 records of gene expression are divided into 161 samples for training and 69 samples for testing. The heart rate variability dataset's 928 records are partitioned into 650 samples of training and 278 samples of testing data.

5. Conclusion

We have proposed a system for predicting risk score for cardiac arrest recurrence. High accuracy was achieved using ensemble classifiers as compared with existing machine learning algorithms available. The approach used HRV features for the prediction of final risk score. Diabetes, blood pressure and gene expression analysis further reinforced the accuracy of the risk score thus predicted. Using the build classifiers, risk score can be predicted for an unknown patient and thus suitable measures can be taken for high-risk patients.

The proposed system uses Wald statistics for risk score calculation. In case of small sample size, Wald statistics may not produce reliable result. Data producing large estimate of coefficient will result in lower Wald statistics due to inflation of standard error and thus a critical feature may be considered as unimportant. To overcome this issue, we would experiment with likelihood ratio tests, which is generally considered to be superior, for weight calculation.

Acknowledgements

We would like to thank Dr G R C Reddy, Director, National Institute of Technology Goa, for providing support to this research work.

References

[1] Fishman G I, Chugh S S, DiMarco J P, Albert C M, Anderson M E, Bonow R O, Buxton A E, Chen P S, Estes M, Jouven X, *et al* 2010 Sudden cardiac death prediction and prevention report from a National Heart, Lung, and Blood Institute and Heart Rhythm Society workshop. *Circulation* 122(22): 2335–2348

[2] Nichol G, Tomas E, Callaway C W, Hedges J, Powell J L, Aufderheide T P, Rea T, Lowe R, Brown T, John D, *et al* 2008 Regional variation in out-of-hospital cardiac arrest incidence and outcome. *J. Am. Med. Assoc.* 300(12): 1423–1431

[3] Chugh S S, Jui J, Gunson K, Stecker E C, John B T, Thompson B, Ilias N, Vickers C, Dogra V, Daya M, *et al* 2004 Current burden of sudden cardiac death: multiple source surveillance versus retrospective death certificate-based review in a large US community. *J. Am. Coll. Cardiol.* 44(6): 1268–1275

[4] De Vreede-Swagemakers J J M, Gorgels A P M, Dubois-Arbouw W I, Van Ree J W, Daemen M J A P, Houben L G E and Wellens H J J 1997 Out-of-hospital cardiac arrest in the 1990s: a population-based study in the Maastricht area on incidence, characteristics and survival. *J. Am. Coll. Cardiol.* 30(6): 1500–1505

[5] Byrne R, Constant O, Smyth Y, Callagy G, Nash P, Daly K and Crowley J 2008 Multiple source surveillance incidence and aetiology of out-of-hospital sudden cardiac death in a rural population in the West of Ireland. *Eur. Heart J.* 29(11): 1418–1423

[6] Hua W, Zhang L F, Wu Y F, Liu X Q, Guo D S, Zhou H L, Gou Z P, Zhao L C, Niu H X, Chen K P, *et al* 2009 Incidence of sudden cardiac death in China: analysis of 4 regional populations. *J. Am. Coll. Cardiol.* 54(12): 1110–1118

[7] Rao B H, Sastry B K S, Chugh S S, Kalavakolanu S, Christopher J, Shangula D, Korabathina R and Raju P K 2012 Contribution of sudden cardiac death to total mortality in India population based study. *Int. J. Cardiol.* 154(2): 163–167

[8] Goldberger A L 2013 *Goldberger's clinical electrocardiography: a simplified approach*, 1st edn. Elsevier Health Science, Philadelphia

[9] Friedlander Y, Siscovick D S, Weinmann S, Austin M A, Psaty B M, Lemaitre R N, Arbogast P, Raghunathan T E and Cobb L A 1998 Family history as a risk factor for primary cardiac arrest. *Circulation* 97(2): 155–160

[10] Klapan E I and Meier P 1958 Nonparametric estimation from incomplete observation. *J. Am. Stat. Assoc.* 53: 457–481

[11] Furukawa T, Rozanski J J, Nogami A, Moroe K, Gosselin A J and Lister J W 1989 Time-dependent risk of and predictors for cardiac arrest recurrence in survivors of out-of-hospital cardiac arrest with chronic coronary artery disease. *Circulation* 80(3): 599–608

[12] Mantel N 1967 The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27(2 p1): 209–220

[13] Glümer C, Carstensen B, Sandbæk A, Lauritzen T, Jørgensen T and Borch-Johnsen K 2004 A danish diabetes risk score for targeted screening—the Inter99 study. *Diabetes Care* 27(3): 727–733

[14] Chen L, Magliano D J, Balkau B, Colagiuri S, Zimmet P Z, Tonkin A M, Mitchell P, Phillips P J and Shaw J E 2010 AUSDRISK: an Australian Type 2 Diabetes Risk Assessment Tool based on demographic, lifestyle and simple anthropometric measures. *Med. J. Aust.* 192(4): 197

[15] Sugioka T, Hayashino Y, Konno S, Kikuchi S and Fukuhara S 2008 Predictive value of self-reported patient information for the identification of lumbar spinal stenosis. *Fam. Pract.* 25(4): 237–244

[16] Kulothungan V, Ramakrishnan R, Subbiah M and Raman R 2014 Risk score estimation of diabetic retinopathy: statistical alternatives using multiple logistic regression. *J. Biometr. Biostat.* 5(5): 211

[17] Wald A 1943 Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Am. Math. Soc.* 54(3): 426–482