Evaluation prediction techniques to achievement an optimal biomedical analysis

Samaher Al-Janabi* and
Muhammed Abaid Mahdi

Department of Computer Science,
Faculty of Science for Women (WSCI),
University of Babylon,
Babylon, Iraq
Email: Samaher@uobabylon.edu.iq
Email: wsci.muhammed.a@uobabylon.edu.iq
*Corresponding author

Abstract: Intelligent analysis of prediction data mining techniques is widely used to support optimising future decision-making in many different fields including healthcare and medical diagnoses. These techniques include Chi-squared Automatic Interaction Detection (CHAID), Exchange Chi-squared Automatic Interaction Detection (ECHAID), Random Forest Regression and Classification (RFRC), Multivariate Adaptive Regression Splines (MARS), and Boosted Tree Classifiers and Regression (BTCR). This paper presents the general properties, summary, advantages, and disadvantages of each one. Most importantly, the analysis depends upon the parameters that have been used for building a prediction model for each one. Besides, classifying those techniques according to their main and secondary parameters is another task. Furthermore, the presence and absence of parameters are also compared in order to identify the better sharing of those parameters among the techniques. As a result, the techniques with no randomness and mathematical basis are the most powerful and fast compared with the others.

Keywords: biomedical analysis; data mining; prediction techniques; healthcare problem; parameters.

Reference to this paper should be made as follows: Al-Janabi, S. and Mahdi, M.A. (2019) ‘Evaluation prediction techniques to achievement an optimal biomedical analysis’, *Int. J. Grid and Utility Computing*, Vol. 10, No. 5, pp.512–527.

Biographical notes: Samaher Al-Janabi received her BSc, MSc and PhD degrees in Computer Science from Science College, University of Babylon, Iraq. Through her years of study, she specialised in the design, implementation and performance measurement and intelligent analysis of huge/bigdata, databases. Her research interests span topics concerning intelligent data analysis, knowledge discovery in databases, soft computing techniques, artificial intelligence, data mining, prediction techniques, mobile services, internet of things and network security. She has published well over 49 scientific papers and authored three books; one on new trends of KDD and one book on soft computing techniques, while, the third intelligent miner of huge medical database. She is a one from five women winner the L’Oreal–UNESCO for Women in Science Levant and Egypt Regional Fellowships 2014. She has gotten Patent 2018. She is a Reviewer of several local and international journals.

Muhammed Abaid Mahdi received his BSc and MSc degrees in Computer Science from Science College, University of Babylon, Babylon, Iraq, and the PhD degree in Information Technology College, from the University of Babylon, Babylon, Iraq. His current research interests include computer networking, network communication, networking, information and communication technology, wireless computing, network simulation, mobility management, social networks and social applications. He gotten award of the best paper (*A Developed Realistic Urban Road Traffic in Erbil City Using Bi-directionally Coupled Simulations*) has published multi scientific papers. He is a Reviewer of several local and international journals.

1 Introduction

Healthcare problem is one of the most important problems in our modern society; not only it has a significant effect on people’s lives, but also it plays a central role in countries’ financial resources. The high quality of healthcare system contributes with other factors in increasing of life expectancy, for example, the average age of death in the United States reached 68.2 in 1950 and increased to 78.7 in 2010 (National Centre for Health Statistics, 2012) while the cost of this care grew rapidly and reached 2.6 trillion dollars in 2010 (National Vital Statistics Reports, 2013). There are some major control factors that have a direct effect on those costs, such as the number of patients and the number of days they will spend in hospitals and unnecessary hospital admissions, which could play a root cause in wasting of resources.

Data Mining (DM) is the core of the Knowledge Discovery in Database (KDD), we can achieve multi-tasking by it as explained in Figure 1.

Predicting is one of the data mining techniques used in many different scientific disciplines to find and analyse historical data and make a prediction by using different techniques such as statistics, data mining and machine learning (Ali, 2012). Statistical methods such as regression can mathematically represent the interaction between different variables under some considerations. Later on, Artificial Intelligence (AI) techniques gained popularity and replaced the statistical model in most applications because they are known to be efficient and less time-consuming in the modelling of complex systems compared to mathematical models such as regression (Yadav and Chandel, 2014).

In general, we can consider prediction as one of the regression methods related to supervised learning, as explained in Figure 2. The main purpose of this paper is to determine the best prediction data mining techniques based on their performance. This pre-processing one of the healthcare datasets, writing a complete algorithm for each prediction techniques as pseudo code and determining the main parameters for each one.

In recent years several researchers have investigated the use of prediction techniques in the healthcare sector. Peng et al. (2011) used machine learning algorithms to reduce unnecessary hospitalisations by predicting how long the patient will stay in the hospital in the next year according to his last year's records. The authors used SVM, Random forest, Regression tree and Boosting Ensemble with HPN 2011 Dataset (HPN, 2011).

Moon et al. (2012) developed the Decision Tree model to discover patterns in smoking behaviours among elders and researched factors by using the CART method based on the National Survey on Drug Use and Health (NSDUH, 2006) in an attempt to decrease heavy smoking habits. The researchers compared the proposed model with the Logistic Regression model by using the accuracy performance measure (Rahman and Hasan, 2011).

Yao et al. (2013) introduced a novel method to predict diseases by a combination of the Random Forest and Multivariate Adaptive Regression Splines on The Wisconsin Diagnostic Breast Cancer data set (WDBC). The researchers found that the RF & MARS method is higher classification accuracy than the RF model, but lower classification accuracy than the MARS model. The accuracy, the sensitivity, the specificity, and the confusion matrixes were used as a basic measure to evaluate the performance of the proposed method (Moon et al., 2012).

Decision Tree C4.5, Naïve Bayesian classifier, and Support Vector Machine (SVM) was used by Khalilinezhad et al (2015) to predict the healthiness of blood donors in Blood Transfusion Organisation (BTO). The results revealed that the SVM was more accurate with lower error compared with the other two algorithms (Khalilinezhad et al., 2015).

The remainder of this paper is as follows. Section 2 shows methods and materials in two sub-parts: first Data Collection and Processing while the second part Prediction Techniques used to Take Decision that include **CHAID** E-CHAID, **RFRC**, MARS, and BTCR. Section 3 shows implementation of prediction techniques and analysis results. Finally, Section 4 describes and discusses the conclusion of this paper, with highlight points for the main hypothesis, limitations, advantages and disadvantages related to a case study.

Figure 1 Basic data mining tasks

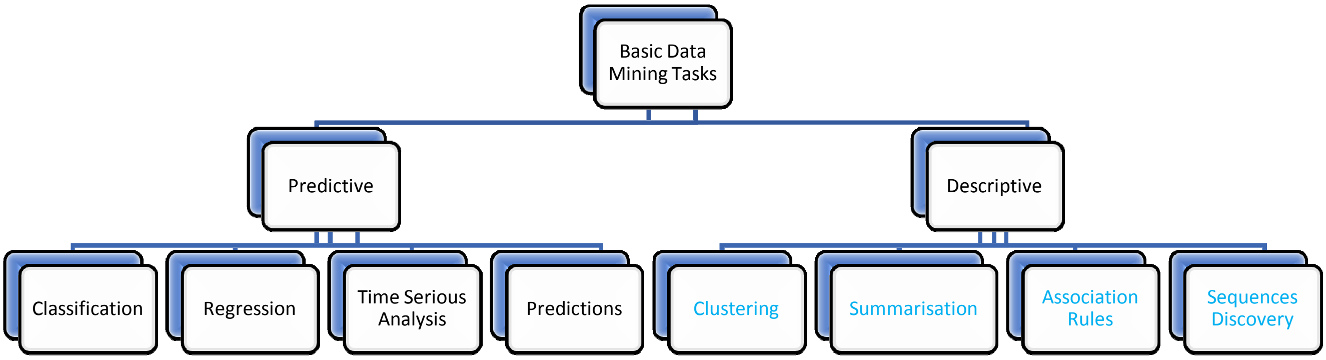
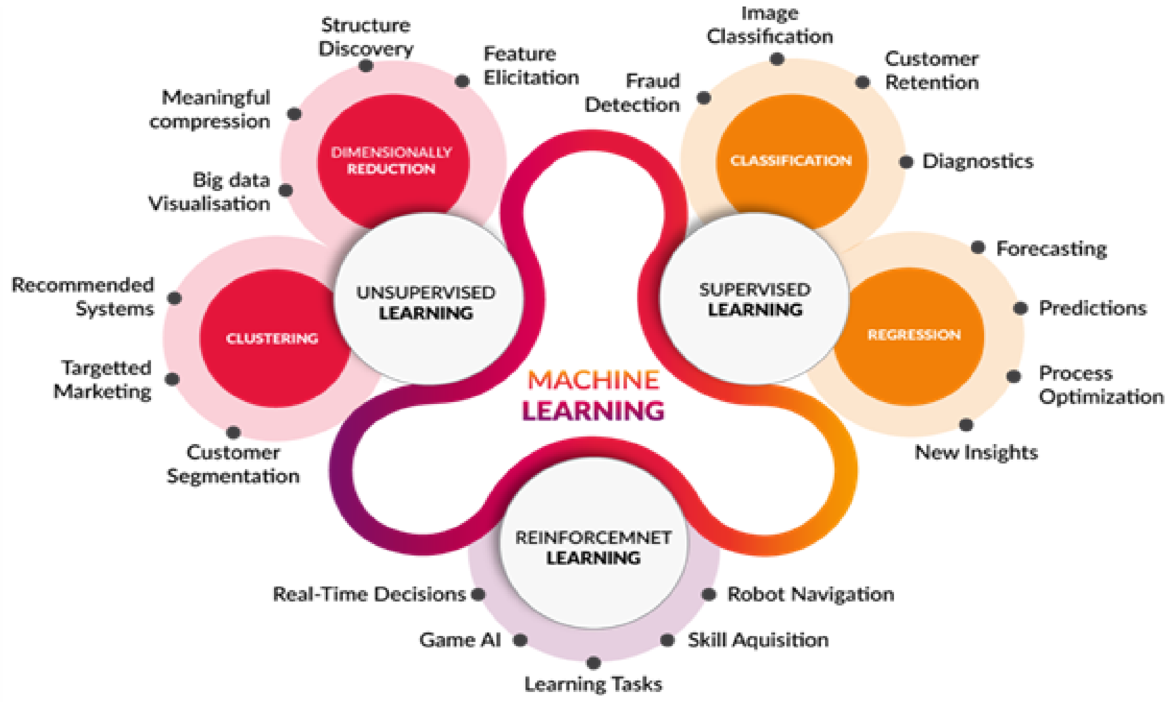


Figure 2 Location of prediction in machine learning map



2 Methods and materials

This section will explain the main steps and the prediction materials used to handle the Echocardiogram data set; in general, the analysis of that prediction material also will show in this section. A list of abbreviations used in this paper is provided in Table 1.

2.1 *Data collection and processing*

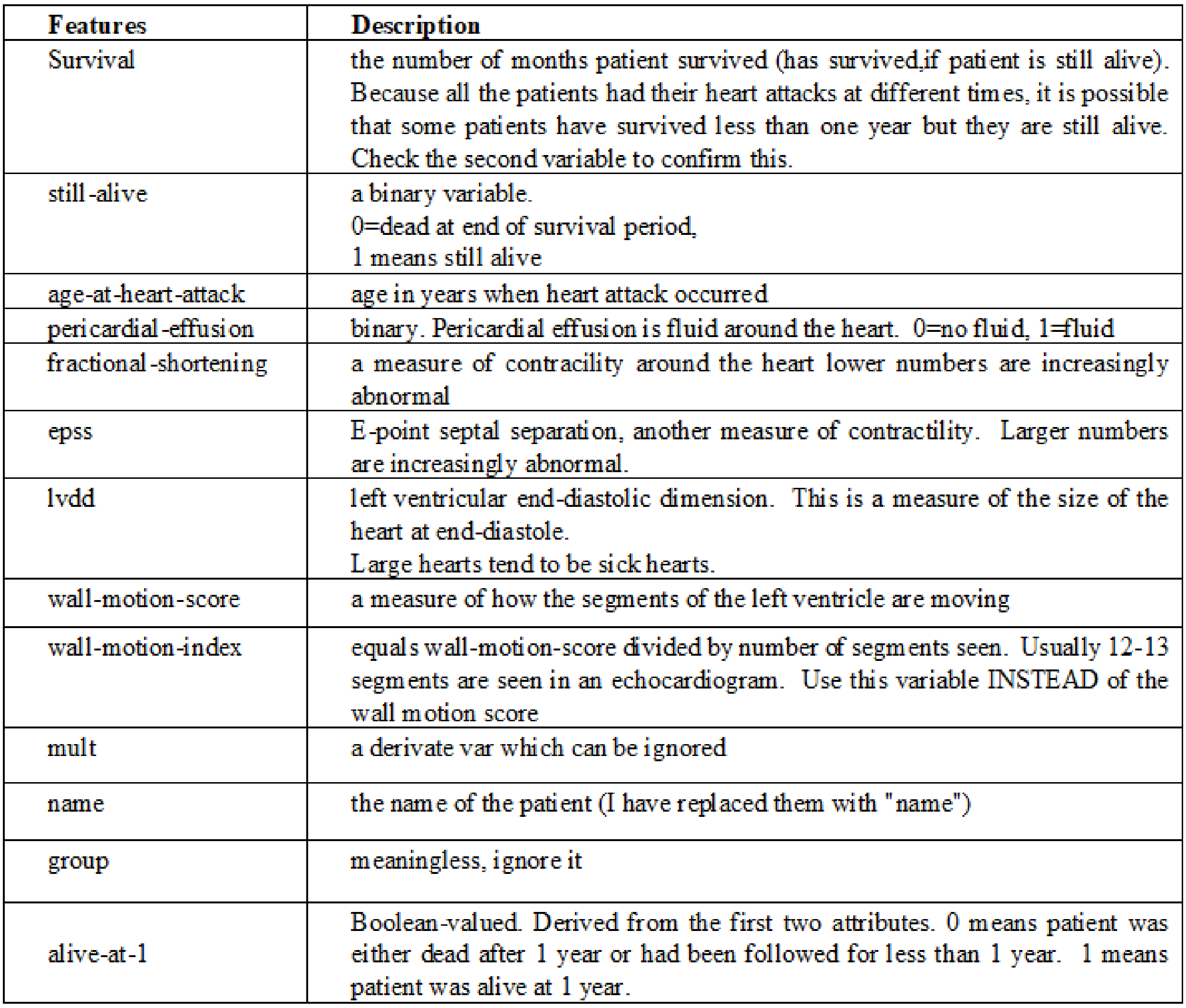
We applied the prediction techniques to a real-data set of (Echocardiogram Data[1]). The metadata of this data set refer to the following information; number of instances: 132, number of attributes: 13 (all numeric-valued), all the patients suffered heart attacks at some points in the past. Some are still alive and some are not. The most difficult part of this problem is correctly predicting that the patient will not survive (http://www.sgi.com/tech/mlc/db/echocardiogram.data). We can explain the main details of the attribute information in Table 2.

- *Age at heart attack*: continuous.
- *Pericardial effusion*: 0, 1
- *Fractional shortening*: continuous.
- *E-point septal_separation*: continuous.
- *Left ventricular end-diastolic dimension*: continuous.
- *Wall-motion-score*: continuous.
- *Wall-motion-index*: continuous.

Table 1 List of abbreviations definitions

Abbreviations	Description
DM	Data Mining
KDD	Knowledge Discovery in Database
NSDUH	National Survey on Drug Use and Health
WDBC	Wisconsin Diagnostic Breast Cancer
SVM	Support Vector Machine
BTO	Blood Transfusion Organisation
BTCR	Boosted Tree Classifiers and Regression
CHAID	CHi-squared Automatic Interaction Detection
E-CHAID	Exhaustive CHAID
MARS	Multivariate Adaptive Regression Splines
RFRC	Random Forest Regression and Classification
P(X)	Prior Probability of X
P(Y)	Prior Probability of Y
X	Interest Variables
W	Case Weight
F	Frequency Weight
try	Randomly Sample of The Predictors
B	Number of Iterations
in	Minimum Number of Samples in Node
M	Number of Mars Terms
K	Karnal Function
C	kernel Function
Y	Target Variable

Table 2 Features and their description



Features	Description
Survival	the number of months patient survived (has survived,if patient is still alive). Because all the patients had their heart attacks at different times, it is possible that some patients have survived less than one year but they are still alive. Check the second variable to confirm this.
still-alive	a binary variable. 0=dead at end of survival period, 1 means still alive
age-at-heart-attack	age in years when heart attack occurred
pericardial-effusion	binary. Pericardial effusion is fluid around the heart. 0=no fluid, 1=fluid
fractional-shortening	a measure of contracility around the heart lower numbers are increasingly abnormal
epss	E-point septal separation, another measure of contractility. Larger numbers are increasingly abnormal.
lvdd	left ventricular end-diastolic dimension. This is a measure of the size of the heart at end-diastole. Large hearts tend to be sick hearts.
wall-motion-score	a measure of how the segments of the left ventricle are moving
wall-motion-index	equals wall-motion-score divided by number of segments seen. Usually 12-13 segments are seen in an echocardiogram. Use this variable INSTEAD of the wall motion score
mult	a derivate var which can be ignored
name	the name of the patient (I have replaced them with "name")
group	meaningless, ignore it
alive-at-1	Boolean-valued. Derived from the first two attributes. 0 means patient was either dead after 1 year or had been followed for less than 1 year. 1 means patient was alive at 1 year.

2.2 *Prediction techniques used to take decision*

In this study, different prediction data mining techniques were used and briefly analysis of their properties are explained.

2.2.1 *Chi-squared automatic interaction detection (CHAID)*

This technique is considered one of the decision tree techniques in which each parent node has more than two children. CHAID needs to transform continuous predictors to categorical ones (Yao et al., 2013), through allowing to multi split of each node this given more chance for each variable to appear. CHAID is suitable for only the nominal or ordinal categorical variables; it needs more time of pre-processing if variables are continuous because it must transform them into ordinal variables. In addition, CHAID requires the user to determine many parameters.

Procedure of CHAID

Input: Echocardiogram data

Output: Predicted attributes.

1 Determined the goal attribute Y.

2 Determined the set of important attributes X.

3 For each variable $x \in X$, do following merging steps:

3.1 For each two adjacent categories ci, $ci+1 \in C$, find a p-value from the following equation:

$$F = \frac{\sum_{l=1}^{I}\sum_{n\in D} w_n f_n I(x_n = i)(\bar{y}_i - \bar{y})^2 / (I-1)}{\sum_{l=1}^{I}\sum_{n\in D} w_n f_n I(x_n = i)(y_n - \bar{y}_i)^2 / (N_f - I)}$$

3.2 If the largest p-value (lowest effect on Y) > α merge (alpha_merge), then merge two adjacent categories of this p-value.

```
4 From all variables X, choose x that has the lowest a p-value (largest effect on Y).
5 If a p-value < α split (alpha_split), then based on classes of attribute x from step 4 dived node n to k child nodes.
      Else A node n called a "terminal node".
6 Return to 4 when stopping condition not satisfy.
      Else go to 7.
7 Tracing the final tree model to find the values of goal attribute Y.
8 Using mean or median to get the value of goal attribute if tracing stop on a terminal node.
9 End procedure
```

2.2.2 Exhaustive Chi-squared automatic interaction detection (E-CHAID)

E-CHAID is similar to CHAID in point the parent node can divide into more than two children which means a generated general tree rather than binary tree. "The main difference is that ECHAID merges more steps. To merge steps is an exhaustive search procedure in merging any similar pair until only a single pair remains and a p-value compare with the previous step rather than with user-specific parameter" (StatSoft, 2010). This technique needs the user to determine fewer parameters compare with CHAID; therefore, this point is considered advantage of that technique.

One of ECHIAD's advantages is that it needs fewer user specific parameters than CHAID: no alpha-level α merge (alpha_merge), or alpha-level split-merge α (alpha split-merge) are needed that leads to more automatic operations.

Procedure of E-CHAID

Input: Echocardiogram data

Output: Predicted attributes.

```
1 Determined the goal attribute Y.
2 Determined the set of important attributes X.
3 For each variable x ∈ X, do following merging steps:
    3.1. For each two adjacent categories ci, ci+1 ∈ C, find a p-value from the following equation:
         $F = \frac{\sum_{l=1}^{l}\sum_{n\in D} w_n f_n I(x_n = i)(\bar{y}_i - \bar{y})^2 / (I - /1)}{\sum_{l=1}^{l}\sum_{n\in D} w_n f_n I(x_n = i)(y_n - \bar{y}_i)^2 / (N_f - I)}$
    3.2. If the largest a p-value (lowest effect on Y) > α merge (alpha_merge), then merge two adjacent categories of this p-value.
    3.3. Repeat steps 3.1 and 3.2 until two categories remain.
4 From all variables X, choose x that has the lowest a p-value (largest effect on Y).
5 If a p-value < α split (alpha_split), then based on classes of attribute x from step 4 dived node n to two child nodes.
      Else A node n called a "terminal node".
6 Return to 4 when stopping condition not satisfy.
      Else go to 7.
7 Tracing the final tree model to find the values of goal attribute Y.
8 Using mean or median to get the value of goal attribute if tracing stop on a terminal node.
9 End procedure
```

2.2.3 Random forest regression and classification (RFRC)

The RFRC is one of the combination algorithms based on the idea take the final decision by apply the higher voting principle on forest of trees (Al-Janabi et al., 2014). RF used the two types of choose randomly; the first when selection number of samples from total number of samples to build tree; while the second random selection when choose the random number of features from total features to build each sub tree. It is simple and easily parallelised (Breiman, 2001; Kursa, 2013).

Procedure of RFRC

Input: Echocardiogram data

Output: Predicted attributes.

```
1 Determined the goal attribute Y.
2 Determined the set of important attributes X.
3 Divided Dataset to n tree samples, one for each tree.
4 For each tree t, do the following steps:
    4.1 For each node n in t, do the following:
        4.1.1 Choose random my set of variables X.
        4.1.2 Select optimal dived of variable x to split node n.
        4.1.3 If satisfy the splitting condition, then divide n.
              Else A node n called a "terminal node".
    4.2 Make pruning for tree t.
    4.3 Calculate the mean of squared residuals of OOB according to the following equation:
5 After Radom Forest of n-tree trees is completed then predicate value of Y by find mean or median of n-tree prediction.
6 End procedure.
```

2.2.4 Multivariate adaptive regression splines (MARS)

The MARS is one of the prediction techniques extraction from divide and conquers principle. It works by dividing the input variables into multi region then determined number of mathematical equation (bias equations) and coefficients.

MARS handle multi-dimension variables (2 to 20 variables) (Friedman, 1990).

It automatically extends to cover nonlinear relation between a dependent variable (*Y*) and an independent variable (*X*). Basis functions are used to specify the relation between *Y* and *X* for each equation. Every two basic functions have a knot (shared point in the decision boundary) that is specified from the data. When the terms of the basic functions are complete, backward steps are needed as a post-pruning to avoid over-fitting (Moon et al., 2012; Statsoft, 2010).

MARS does not have a tree like CART (i.e., for more details of CART see (Al-Janabi, S. and Al-Shourbaji, 2016)) or CHAID but a series of equations which perform regression tasks, thus, it depends totally on mathematical functions having the strength of mathematics for finding the optimal solution. MARS has many interesting features. No user specific parameters are needed making it more flexible because it adapts to data. Also, the variables do not need a transformation that eliminates pre-processing steps which in turn leads to less computational time. Variables are automatically selected by MARS. Another interesting feature is that MARS has the ability to handle more than one target variable (*Y*). Despite being of the slower model building compared with recursive partitioning, MARS has a fast prediction with new unseen data. Sometimes, it suffers from a discontinuity in sub-region boundaries that may affect the accuracy (Moon et al., 2012; Statsoft, 2010; Friedman, 1990).

Procedure of MARS

```
Input: Echocardiogram data
Output: Predicted attributes.
    1 Determined the goal attribute $Y$.
    2 Determined the set of important attributes $X$.
    3 Building Model by following steps:
        3.1 While the complexity of Model < $M$, do the following steps:
            3.1.1 For each variable $x$, do following
                  $(x-t)_+ = \begin{Bmatrix} x-t & x>t \\ 0 & \text{otherwise} \end{Bmatrix}$
                A. For each Knot of
                   variable $x$, Test each Knot according to the equation:
                B. Choose Knot for variable $x$, which decrease prediction error.
            3.1.2 Add new basis function from variable $x$ with a knot to the Model.
        3.2 IF complexity of Model >= $M$, then Stop building Model.
    4 For each basis function in the Model make pruning as following steps:
        4.1 Calculate Generalised Cross Validation error according to the equation:
            $GCV = \frac{\sum_{i=1}^{N}(y_i - f(x_i))^2}{\left(1\frac{C}{N}\right)^2}$
            where $N$: number of samples, $C = 1 + cd$
        4.2 Remove function with a high Generalised Cross Validation error.
    5 After Model is completed, predicted the value of $Y$ by using this equation:
        $y = f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m H_{km}(x_{v(\text{km})})$
    6 End procedure.
```

2.2.5 *Boosted tree classifiers and regression (BTCR)*

The BTCR like RFRC in generated sequence of binary tree based on apply forward strategy but it differ from RFRC in point it does not depend on randomness in sampling or select variables (i.e., BTCR is consider a finite loop of CART). "Each iteration tree is built to predicate residual of the previous tree. It prevents the tree from growing without control. In each epoch, the size of the current tree is detected to a fixed value" (Friedman, 1999; Jun, 2013; Rand. 2005; StatSoft, 2010; Elith el al., 2008).

BTCR avoids over-fitting by using limitation on the number of iterations (Jun, 2013). It performs better than RFRC with enough number of samples and handles many types of target variables (*Y*). "There is also no need for data transformation or elimination of outliers and it can automatically handle interaction effects between predictors" (Elith et al., 2008).

The limit of a tree size affects BTCR prediction. BTCR will be poor if choosing the size of the tree is incorrect "because it does not handle the interaction between the variables. Also, poor prediction can happen with a small number of samples for generalisation" (Jun, 2013).

Procedure of BTCR

```
Input: Echocardiogram Data, $J$, $B$.
Output: Predicted attributes.
    1 Determined the goal attribute $Y$.
    2 Determined the set of important attributes $X$.
    3 For $i = 1$ to $B$, do following step for each tree $t_i$
        3.1 Build tree $t_i$ on current data sample ($r_{i1}$, $x$) by following steps:
            3.1.1 For each node $n$ in tree $t_i$, choose best variable $x$ as a splitter.
            3.1.2 Test split condition is correct,
                  divided node n into two child nodes.
                  Else If a number of terminal nodes < $J$,
                  $A$ node $n$ called a "terminal node".
                    Else Stop building tree, go to 3.2
        3.2 For $j = 1$ to $N$, ($N$: number of data sample)
```

```
            Compute residuals according to following equation:
            $r_{ij} = y_j - f(x_j)$
        3.3 Data sample for the next iteration will be $(r_i, x)$.
    4   After Boosted of $B$ trees is complete, predicate values of $Y$ by find mean or median of $B$ prediction trees.
    5   End Procedure.
```

Each prediction technique has a number of parameters, they can be primary factors in building the model of prediction or they can be secondary ones which contribute with the primary parameters to have optimal results. The advantages and disadvantages, as well as main and secondary parameters, are summarised in Table 3 (Al-Janabi, 2015). While Table 4 Comparison of the steps of the main techniques of prediction (Al-Janabi, 2015).

Table 3 Comparison of main techniques of prediction (Al-Janabi, 2015)

Main parameters	Disadvantages	Advantages	Techniques
Y , X , *W* , *F* , Alpha_Merge, Alpha_Split-Merge, Alpha_Split	-Variables transformation - Three user specific parameters . -Time for merge steps	- Allow multi split. Handle market segmentation	**CHAID**
Y , X , *w* , *f* , alpha_split	-Variables transformation - Tow user specific parameters . -Time for merge steps	-Allow multi-split. - handle market segmentation -Needs less user specific parameters than chaid	**E-CHAID**
Y , X , *mtry* , *ntree*	Required time from hours to even days of computation, especially for larger sets.	- Good accuracy - Robust to outliers and noise. - Faster than bagging or boosting. -Simple and easily parallelized.	**RFRC**
Y , X	-Discontinuity in subregion boundaries that effect on accuracy. -Need backward steps to fix over fitting.	- No user specific parameters. - No variable transformation - More flexible – Automatic variable selection – Fast prediction.	**MARS**
Y , X , *J* , *B* , *nmin*	- Poor prediction with incorrect tree size limit and with little number of samples for generalization	- Limit number of iterations. - Handle many type of target variable y . - No need for data transformation or elimination of outliers.	**BTCR**

Table 4 Comparison of the steps of the main techniques of prediction (Al-Janabi, 2015)

BTCR	MARS	RFRC	ECHAID	CHAID	Step
✓	✓	✓	✓	✓	Specify X , Y
×	×	×	✓	✓	Variable transfor_mation
×	×	×	✓	✓	Merge categories
×	×	✓	×	×	Choose random variable subset
✓	✓	✓	✓	✓	Find split variable
✓	×	✓	✓	✓	Find split condition
✓	×	✓	✓	✓	Split parent node
✓	×	✓	✓	✓	Check splitting rule
✓	✓	×	✓	✓	Pruning
✓	×	✓	×	×	Comparing Trees
×	×	×	×	×	Caculate prior pro. P(X)P(Y)
×	×	×	×	×	Find optimal hyperline splitter

3 Implementation the prediction techniques and analysis results

The data set (echocardiogram data set) was analysed using CHAID, E-CHAID, RFRC, MARS, and BTCR techniques. The performance of these prediction techniques was investigated and calculated using several measures to evaluate the difference between the expected and actual values.

Table 5 Described the natural of each feature in data set

Number	*Variable Categories*	*Class*	*Type*	*Missing*	*Rows*
1	age at heart attack	Predictor	Continuous	5	
2	still-alive	Predictor	Continuous	0	
3	fractional-shortening	Predictor	Continuous	7	51
4	E-point septal separation	Predictor	Continuous	14	82
5	left ventricular end-D.D.	Predictor	Continuous	10	94
6	wall-motion-score	Predictor	Continuous	3	45
7	wall-motion-index	Predictor	Continuous	1	59
8	alive-at-1	Target	Continuous	0	

Table 5 shows the natural of each feature in data set from points “variable Categories, Class, Type and Missing, number of samples that contain this feature”.

Table 6 describes the important for each features order from higher to lower, here as appear wall-motion-index is consider more important feature.

Table 6 Important of each features

Variable	*Importance*
wall-motion-index	100.000
E-point septal separation	96.408
fractional-shortening	88.945
age at heart attack	88.512
wall-motion-score	87.194
left ventricular end-diastolic dimension	83.454
still alive	14.665

3.1 *Analysis the performance of CHAR*

Table 7 summarises the main parameters required to implement CHAR where this model shows the variables split into dependent and independent; it determined Maximum Tree Depth, Minimum Cases in Parent Node and Minimum Cases in Child Node after that show results. More details about it see Figure 3.

Figure 3 Tree diagram of CHAID

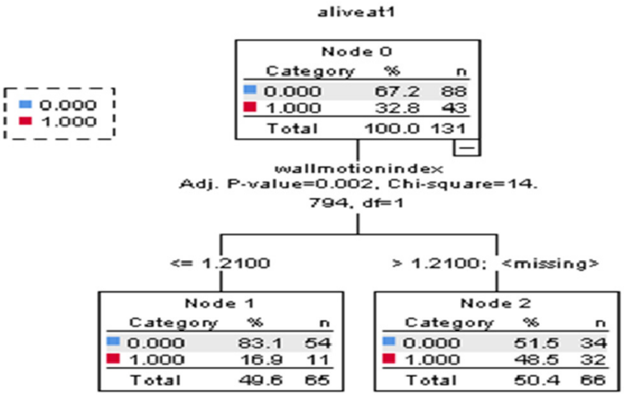


Table 7 Summary of CHAR model

Specifications	Growing method	CHAID
	Dependent variable	aliveat1
	Independent variables	ageatheartattack, stillalive, fractionalshortening, Epointseptalseparation, leftventricularenddiastolicdimension, wallmotionscore, wallmotionindex
	Validation	None
	Maximum tree depth	3
	Minimum cases in parent node	100
	Minimum cases in child node	50
Results	Independent variables included	wall_motion_index
	Number of nodes	3
	Number of terminal nodes	2
	Depth	1

Table 8 Risk of estimate and std. error to CHAID model

Estimate	*Std. error*
.328	.041

Table 9 Confession matrix related to CHAID model

Observed	*Predicted*		
	0	*1*	*Per cent correct*
0	88	0	100.0%
1	43	0	0.0%
Overall percentage	100.0%	0.0%	67.2%

3.2 Analysis the performance of E-CHAR

Table 10 Summary of E-CHAR model

Specifications	Growing method	Exhaustive CHAID
	Dependent variable	aliveat1
	Independent variables	ageatheartattack, stillalive, fractionalshortening, Epointseptalseparation, leftventricularenddiastolicdimension, wallmotionscore, wallmotionindex
	Validation	None
	Maximum tree depth	3
	Minimum cases in parent node	100
	Minimum cases in child node	50
Results	Independent variables included	leftventricularenddiastolicdimension
	Number of nodes	3
	Number of terminal nodes	2
	Depth	1

Table 11 Risk of estimate and std. error to E-CHAID model

Estimate	*Std. error*
.099	.026

Table 12 Confession matrix related to E-CHAID

Observed	*Predicted*		
	0	*1*	*Per cent correct*
0	75	13	85.2%
1	0	43	100.0%
Overall percentage	57.3%	42.7%	90.1%

3.3 Analysis of the performance of RFRC

Table 13 Summary of random forest model

Actual Class	*Predicted Class*		*Actual Total*
	0	1	
0	64.00000	24.00000	88.00000
1	2.00000	41.00000	43.000000
Pred. Tot.	66.00000	65.00000	131.00000
Correct	0,72727	0.95349	
Success Ind.	0.05552	0.62524	
Tot. Correct	0.80153		

Table 14 Counts and % correct by predicted and actual class based on random forest model

By predicted class				*By actual class*		
Class	*N*	*Wgt N*	*% Correct*	*N*	*Wgt N*	*% Correct*
0	66	66.00	0.96970	88	88.00	0.72727
1	65	65.00	0.63077	43	43.00	0.95349

Table 15 Confession matrix related to RFC

Actual class	*Total class*	*Per cent correct*	*Predicted Classes*	
			0 *N = 19*	*1* *N = 112*
0	88	15.91%	**14**	74
1	43	88.37%	5	**38**
Total:	131			
Average:		52.14%		
Overall % correct:		39.69%		
Measures				
Specificity		15.91%		
Sensitivity/Recall		88.37%		
Precision		33.93%		
F1 statistic		49.03%		

3.4 Analysis of the performance of MARS

The basis function of MARS is:

BF1 = (e_point_septal_separation in (“22”, “16”, “23”, “11”, “20”, “17”, “15”, “12”, “19”, “12.733”, “?”, “17.2”, “23.6”, “21.3”, “14”, “14.8”, “24.6”, “16.4”, “11.4”, “12.2”, “21.7”, “19.4”, “19.2”, “8.7”, “11.3”, “4.8”, “8.5”, “28.9”, “6.9”, “40”, “7.6”, “28.6”, “12.9”));

BF3 = (wall_motion_index in (“1”, “1.7”, “1.875”, “1.14”, “1.19”, “2”, “1.38”, “1.5”, “1.11”, “1.667”, “1.56”, “1.67”, “1.222”, “1.3”, “1.08”, “1.167”, “1.05”, “1.39”, “1.18”, “1.1”, “1.367”, “2.39”, “1.09”, “1.83”, “1.27”, “1.06”, “1.23”, “2.2”, “1.95”, “1.2”, “1.375”, “1.73”, “1.21”, “1.409”));

BF5 = (e_point_septal_separation in (“12.062”, “22”, “16”, “23”, “11”, “17”, “15”, “12”, “?”, “17.2”, “13.1”, “23.6”, “21.3”, “14”, “14.8”, “24.6”, “18.6”, “9.8”, “16.4”, “11.4”, “16.1”, “12.2”, “21.7”, “19.4”, “25”, “8.7”, “4.8”, “8.5”, “28.9”, “6.9”, “40”));

BF7 = (wall_motion_score in (“14”, “16”, “18”, “12”, “22.5”, “15.5”, “11.67”, “24”, “27”, “19.5”, “13.83”, “7.5”, “10”, “2”, “6”, “13”, “5”, “21.5”, “15”, “11”, “22”, “17”, “23”, “39”, “12.33”, “10.5”, “16.67”, “17.83”, “5.5”, “16.5”, “21”, “28”, “11.5”, “13.5”, “12.67”, “12.5”, “26.08”, “18.16”, “19”, “14.5”));

BF9 = (wall_motion_index in (“1”, “1.7”, “1.875”, “1.14”, “1.19”, “2”, “1.333”, “1.38”, “1.5”, “1.11”, “1.667”, “1.56”, “1.17”, “1.67”, “1.222”, “1.3”, “1.25”, “?”, “1.08”, “1.167”, “1.05”, “1.39”, “1.18”, “1.1”, “2.39”, “1.09”, “1.83”, “1.27”, “1.06”, “1.42”, “1.23”, “2.2”, “1.95”, “1.2”, “1.375”, “1.73”, “1.21”, “1.409”));

BF13 = (wall_motion_score in (“14”, “16”, “18”, “12”, “22.5”, “15.5”, “11.67”, “24”, “8”, “27”, “19.5”, “13.83”, “7.5”, “10”, “2”, “?”, “6”, “13”, “21.5”, “15”, “11”, “22”, “17”, “39”, “10.5”, “17.83”, “5.5”, “13.67”, “16.5”, “21”, “11.5”, “13.5”, “12.67”, “15.67”, “12.5”, “26.08”, “18.16”, “19”, “14.5”));

BF15 = (e_point_septal_separation in (“9”, “6”, “4”, “5”, “31”, “8”, “0”, “13”, “10”, “12.063”, “20”, “19”, “12.733”, “5.9”, “7”, “?”, “4.2”, “17.2”, “5.12”, “9.3”, “4.7”, “17.5”, “9.4”, “15.6”, “18.6”, “9.8”, “11.9”, “10.3”, “13.2”, “9.7”, “8.8”, “10.2”, “7.5”, “30.1”, “17.9”, “7.1”, “16.8”, “19.2”, “5.5”, “11.3”, “6.6”, “9.1”, “16.5”, “5.6”, “11.8”, “14.3”, “7.6”, “12.1”, “13.6”, “9.2”, “28.6”, “19.1”, “6.8”, “25.5”, “12.9”));

Y = 0.883366 + 1.00633 * BF1 - 0.457886 * BF3 - 0.696426 * BF5- 0.493822 * BF7 - 0.519786 * BF9 + 0.30042 * BF13 + 0.273484 * BF15.

MODEL ALIVE_AT_1 = BF1 BF3 BF5 BF7 BF9 BF13 BF15;

3.5 Analysis of the performance of BTCR

Table 16 Confession matrix related to MARS model

Actual class	*Total class*	*Per cent correct*	*Predicted classes*	
			0 *N = 87*	*1* *N = 44*
0	88	97.73%	**86**	2
1	43	97.67%	1	**42**
Total:	131			
Average:		97.70%		
Overall % correct:		97.71%		
Measures				
Specificity		97.73%		
Sensitivity/Recall		97.67%		
Precision		95.45%		
F1 statistic		96.55%		

Table 17 Confession matrix related to BTCR

Actual class	*Total class*	*Per cent correct*	*Predicted classes*	
			0 *N = 20*	*1* *N = 1*
0	12	100.00%	**12**	0
1	9	11.11%	8	**1**
Total:	21			
Average:		55.56%		
Overall % correct:		61.90%		
Specificity		100.00%		
Sensitivity/Recall		11.11%		
Precision		100.00%		
F1 statistic		20.00%		

3.6 Comparing the results of the techniques

In this article, several prediction techniques were applied on echocardiogram data set to determine the effective predictor that would potentially give more accurate results for biomedical analysis. The performance of these prediction techniques was investigated and calculated using several measures to evaluate the difference between the expected and actual values. As explained in Table 18 and Figure 13.

Table 18 Compare the accuracy among the predictors

Measures	*BTCR*	*MARS*	*RFCR*	*E-CHAID*	*CHAID*
Specificity	100.00%	97.73%	15.91%	78.72%	46.64%
Sensitivity/ Recall	11.11%	97.67%	88.37%	32.78%	28.98%
Precision	100.00%	95.45%	33.93%	56.26%	56.26 %
F1 statistic	20.00%	96.55%	49.03%	15.0%	55.0%

Figure 4 Tree diagram of E-CHAID

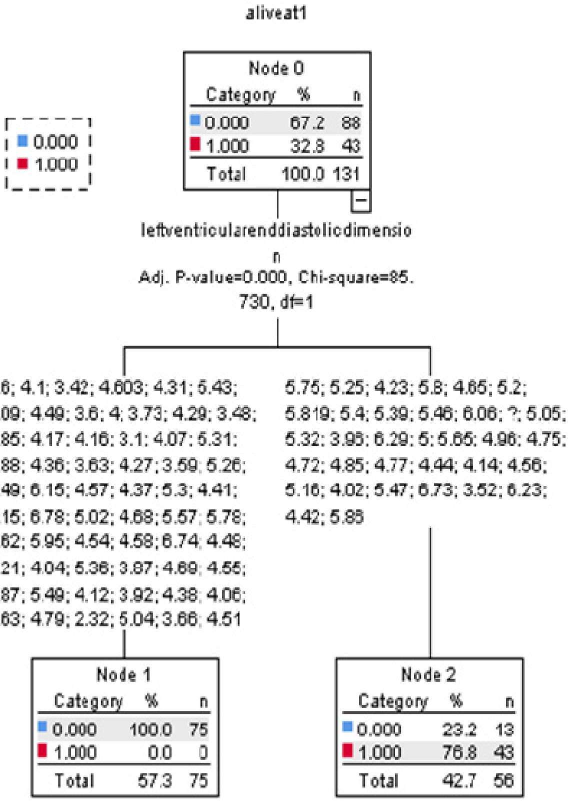


Figure 5 Odds graph of RFC to RESPONSE- OOB and NON-RESPONSE- OOB

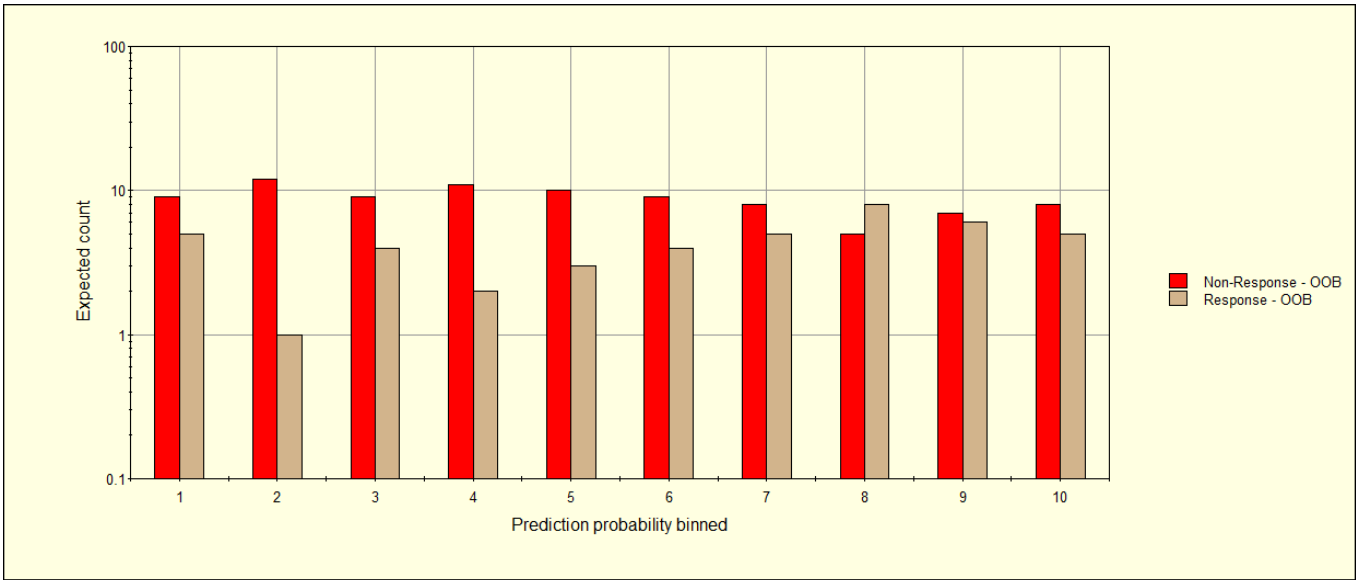


Figure 6 Odds Graph of MARS to RESPONSE- OOB and NON-RESPONSE- OOB

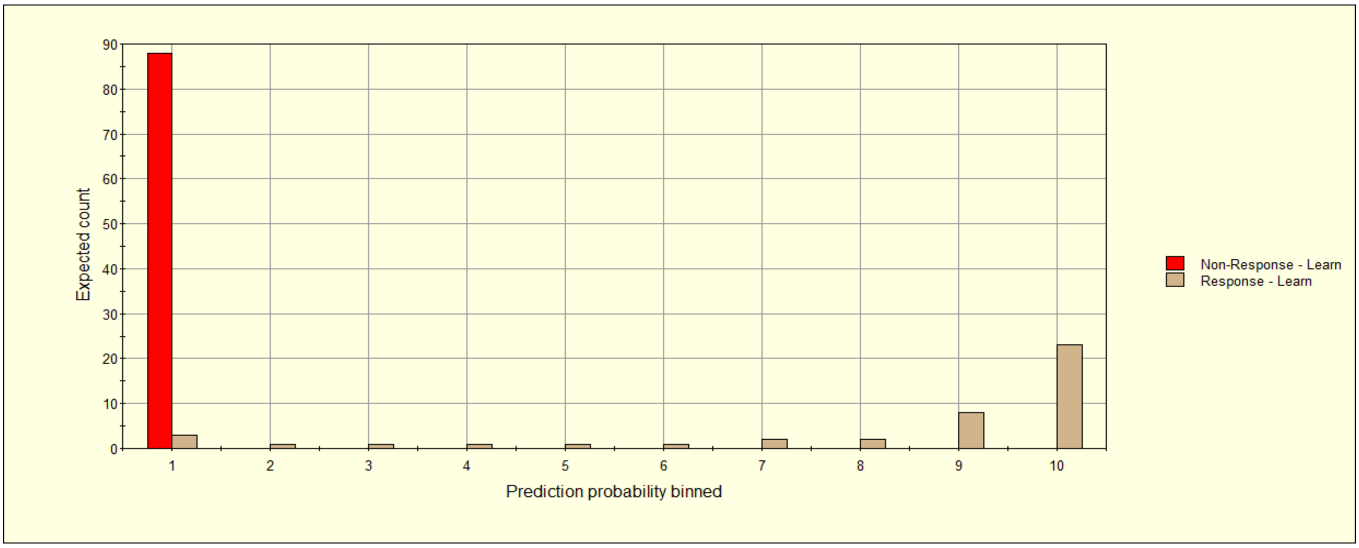


Figure 7 Odds Graph of BTCR to RESPONSE-OBB and NON-RESPONSE-OBB

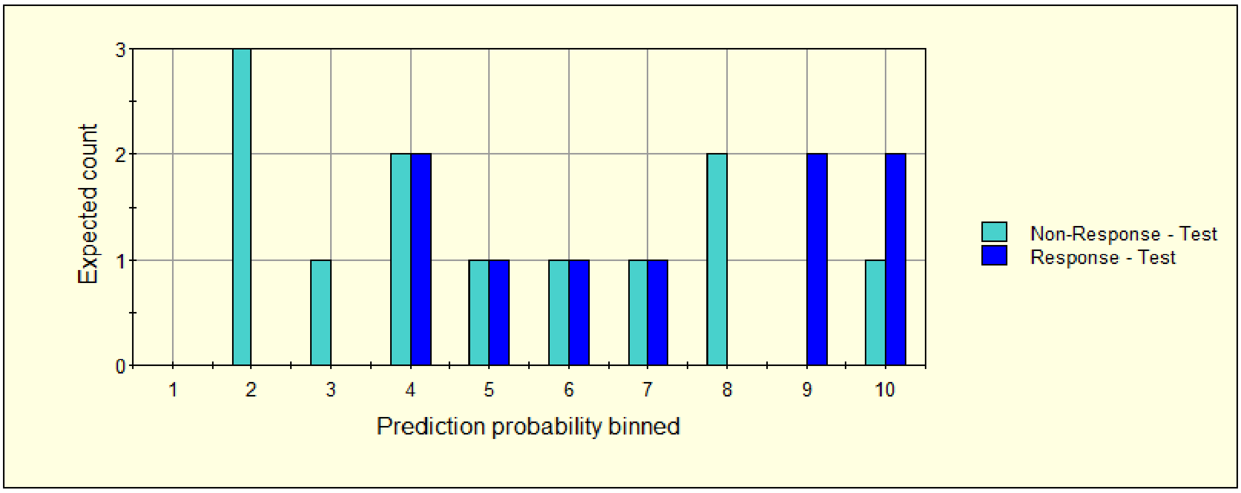


Figure 8 The optimal model based on BTCR

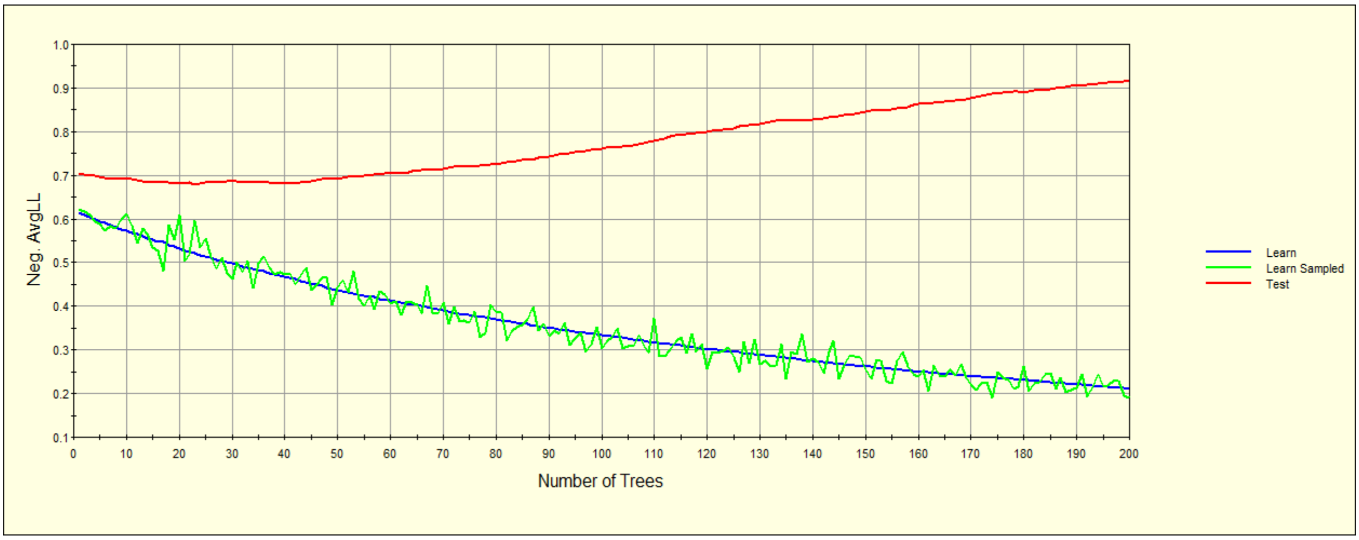


Figure 9 Gain and ROC of the optimal model

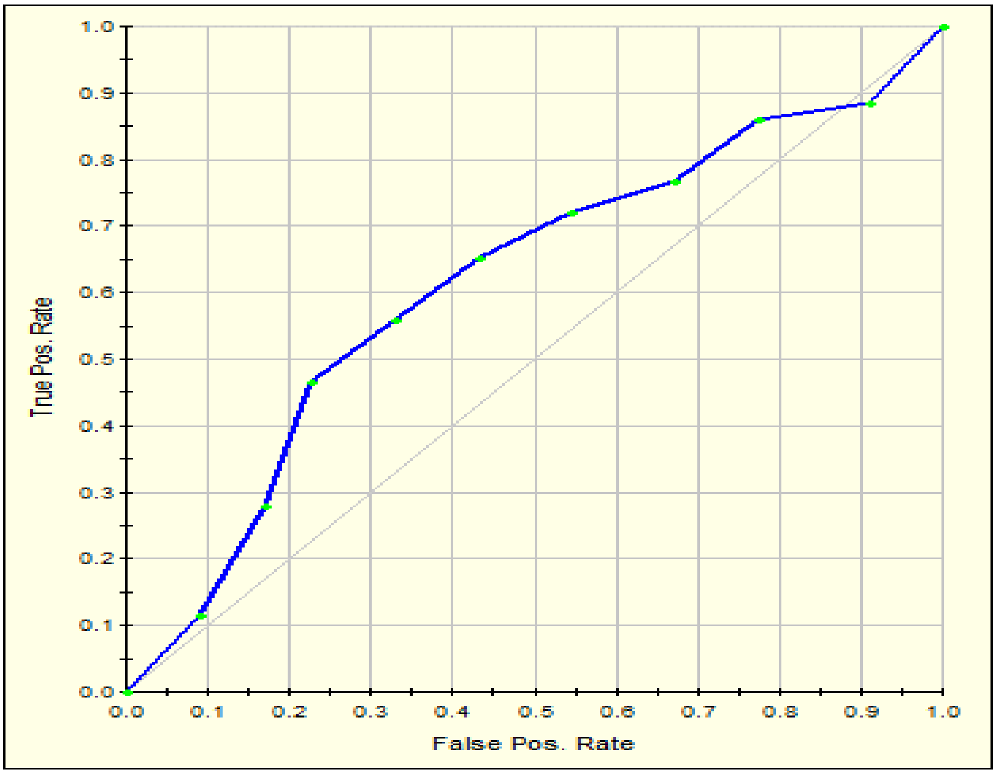


Figure 10 Distributer dataset based on the optimal model (BTCR)

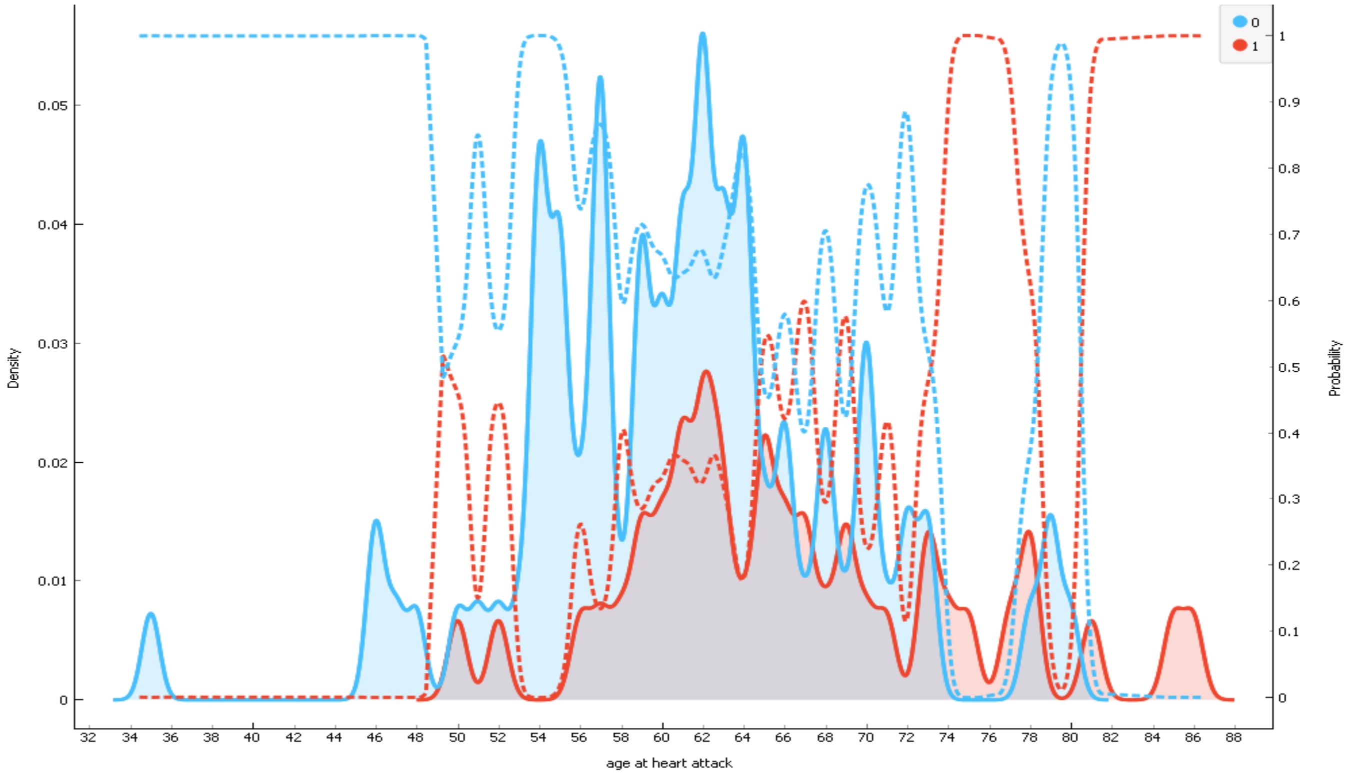


Figure 11 Distributer data set based on best three features in BTCR model

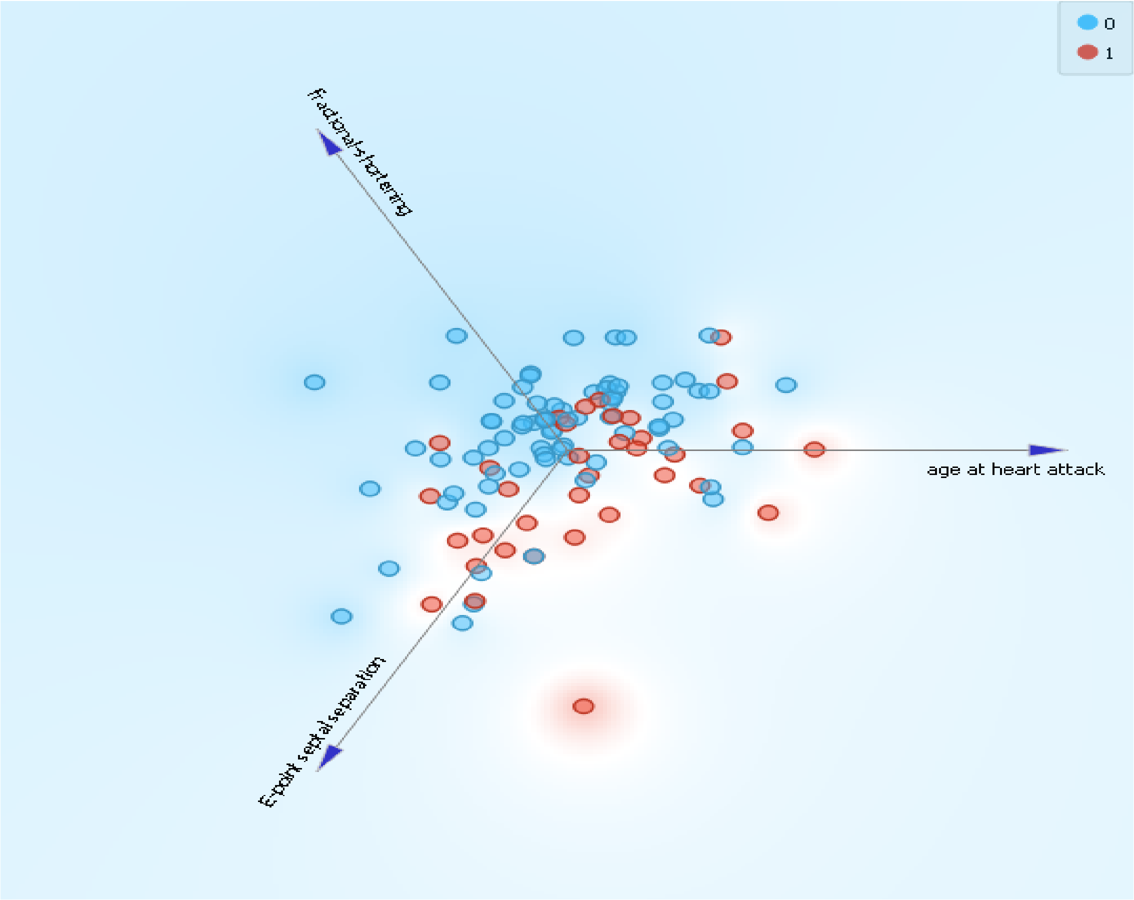


Figure 12 Distributer testing data set based on the optimal BTCR model

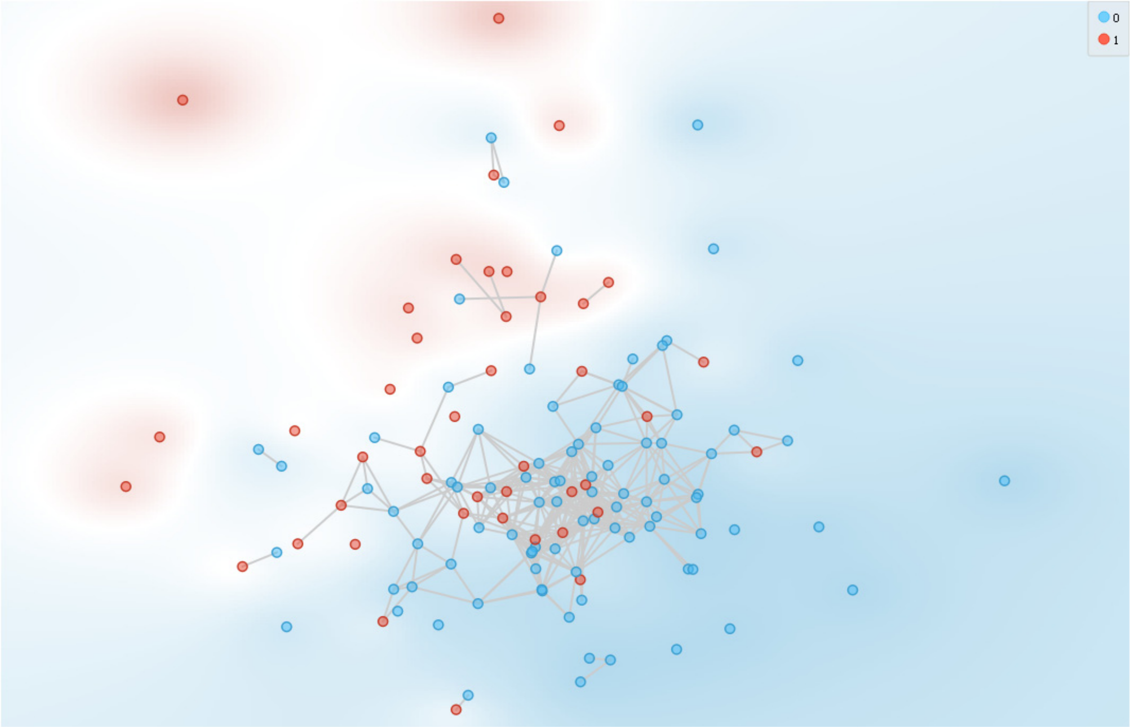


Figure 13 Comparison among the prediction techniques based on evaluation measures

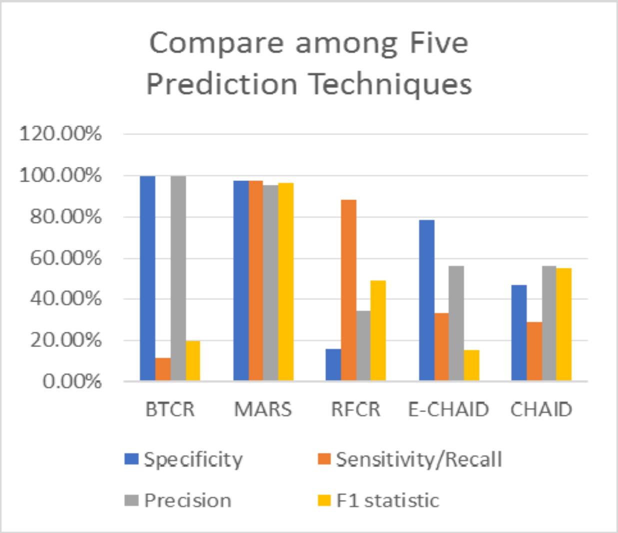
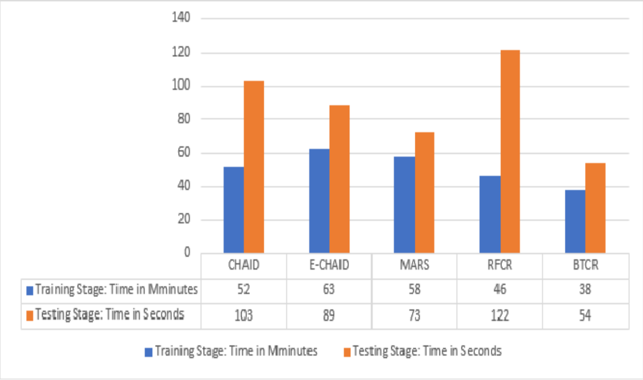


Figure 14 Time required in training and testing stage for each prediction technique



As can be seen in Figure 13, the BTCR predictor is capable of both successfully and accurately forecasting future values.

As results, Table 19 shows BTCR is implemented in less time compared with other prediction techniques, where total time in training stage is 38 minutes while in testing stage it is 54 seconds. Table 19 explains the compare time among the prediction techniques also given in Figure 10.

Table 19 Compare the time among the predictors

Predicator	*Time in minutes: Training phase*	*Time in seconds: Testing phase*
CHAID	52	103
E-CHAID	63	89
MARS	58	73
RFCR	46	122
BTCR	38	54

4 Conclusion

As a result, prediction techniques are powerful tools for solving healthcare problems. The analysis of these techniques refers to some observations. Some of techniques do not depend on randomisation like BTCR, which is better. On the other hand, another technique is more powerful and faster with a mathematical basis like MARS. It gives an optimal solution because of utilising features of mathematics such as linear combination, simplification, derivatives, and integration. Also, from the analysis, we find similar parameters shared among all the prediction techniques like the target variable *Y* and interest variable *X* which are specified as a higher priority. But for this problem, BTCR gives best results comparing with MARS due to the nature of data as explained in obtained results.

References

Al-Janabi, S. (2015) 'A novel agent-DKGBM predictor for business intelligence and analytics toward enterprise data discovery', *Journal of Babylon University/Pure and Applied Sciences*/, Vol. 23, No. 2, pp.482–507.

Al-Janabi, S. and Al-Shourbaji, I., (2016) 'A study of cyber security awareness in educational environment in the middle east', *Journal of Information and Knowledge Management*, Vol. 15, No. 1, p.1650007. doi: 10.1142/S0219649216500076.

Ali, S.H. (2012) 'A novel tool (FP-KC): for handle the three main dimensions reduction and association rule mining', *Proceedings of the 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, Sousse, pp.951–961. doi: 10.1109/SETIT.2012.6482042.

Al-Janabi, S., Patel, A., Atlawi, H., Kalajdzic, K. and Shourbaji, I.A. (2014) 'Empirical rapid and accurate prediction model for data mining tasks in cloud computing environments', *Proceedings of the International Congress on Technology, Communication and Knowledge (ICTCK)*, Mashhad, pp.1–8. doi: 10.1109/ICTCK.2014.7033495.

Breiman, L. (2001) 'Random forest', *Machine Learning*, Vol. 45, pp.5–32.

Elith, J., Leathwick, J.R. and Hastie, T. (2008) 'A working guide to boosted regression trees', *Journal of Animal Ecology*.

Friedman, J.H. (1990) '*Multivariate Adaptive Regression Splines*, Technical Report.

Friedman, J.H. (1999) 'Greedy function approximation: a gradient boosting machine', *The Annals of Statistics*, pp.1189–1232.

HPN (2011) *The heritage health prize competition*. Available online at: http://www.heritagehealthprize.com

Jun, S.H. (2013) *Boosted regression trees and random forests*, Statistical Consulting Report.

Khalilinezhad, M., Minaei, B., Vernazza, G. and Dellepiane, S. (2015) 'Prediction of healthy blood with data mining classification by using Decision Tree, Naive Bayesian and SVM approach', *Proceedings of the 6th International Conference on Graphic and Image Processing (ICGIP'2014) International Society for Optics and Photonics*, pp.94432G–94432G.

Kursa, M.B. (2013) 'Robustness of the Random Forest based gene selection methods, *BMC Bioinformatics*, Vol. 15, No. 8, doi: 10.1186/1471-2105-15-8.

Moon, S.S., Kang, S-Y., Jitpitaklert, W. and Kim, S.B. (2012) 'Decision tree models for characterizing smoking patterns of older adults', *Expert Systems with Applications*, Elsevier, Vol. 39, No. 1, pp.445–451.

National Centre for Health Statistics (2012) *Health*, United States, With Special Feature on Emergency Care, p.342.

National Vital Statistics Reports (2013) Vol. 61, No. 4, May 8, p. 81, Table 8.

Peng, X., Wu, W. and Xu, J. (2011) 'Leveraging machine learning in improving healthcare', *Association for the Advancement of Artificial Intelligence*.

Rahman, R.M. and Hasan, F.R.M. (2011) 'Using and comparing different decision tree classification techniques for mining ICDDR, B hospital surveillance data', *Expert Systems with Applications*, Elsevier, Vol. 38, No. 9, pp.11421–11436.

Rand, M.S. (2005) 'Boosted regression (boosting): an introductory tutorial and a Stata plugin', *The Stata Journal*, Vol. 5, No. 3, pp.330–354.

StatSoft (2010) *Electronic statistics textbook*. Available online at: http://www.statsoft.com/textbook

Yadav, A.K. and Chandel, S.S. (2014) 'Solar radiation prediction using artificial neural network techniques: a review', *Renewable and Sustainable Energy Reviews*, Vol. 33, pp.772–781.

Yao, D., Yang, J. and Zhan, X. (2013) 'A novel method for disease prediction: hybrid of random forest and multivariate adaptive regression splines, *Journal of Computers*, Vol. 8, No. 1, pp.170–177.

Note

1 http://www.sgi.com/tech/mlc/db/echocardiogram.data