Performance Evaluation of Data Mining Algorithm on Electronic Health Record of Diabetic Patients

DrPrakash Kuppuswamy[1], DrRajan John[2], DrShanmugasundaram Marappan[3]

Computer Networks &Engg, Dept., College of CS & IS, Jazan University, Saudi Arabia
Department of Computer Science, College of CS & IS, Jazan University, Saudi Arabia
Department of Computer Science, College of CS & IS, Jazan University, Saudi Arabia
Corresponding Author: Dr Prakash Kuppuswamy

Abstract: *Data mining is a process of finding interesting patterns from large databases. One of the important application areas of data mining is health care sector. In healthcare, it is not only providing useful information to healthcare professionals. It also provides for health insurance companies. Using these techniques many diseases can be predicated at an earlier stage which gives betterment life for human being. In our research work, we have collected Electronic Health Record (EHR) database for a disease of diabetic patients. Every day, the volume of health care data is increasing. Using data mining techniques extract the knowledge from this enormous database efficiently. Many algorithms are available in data mining. We have used classification algorithms such as One R ,Zero R,J48,random forest and LDA. The performance evaluation of classifiers can be analyzed through confusion matrix and in terms of precision, recall and error rate.*
Keywords: *Data mining, Electronic Health Record (EHR), diabetes, Classification algorithms, Zero R, OneR,J48,decision stump and Linear Discriminate Analysis (LDA)*

Date of Submission: 25-02-2019   Date of acceptance: 18-03-2019

I.   Introduction

Health care industry generates large amounts of complex data's such as patient history, hospital resources, electronic records, information about medical devices etc. These data's serves as a key resource to process and analyse for knowledge extraction that enables the decision making and to save cost. Research using data mining techniques have been applied in diagnosis of various diseases such as cardiovascular diseases, AIDS, asthma and diabetes [1]. Diabetes is one of the major health problems of all over the world [2]. Diabetes is a disease that occurs when the insulin production in the body is inadequate or the body is unable to use the produced insulin in a proper manner, as a result, this leads to high blood glucose. The body cells break down the food into glucose and this glucose needs to be transported to all the cells of the body. The insulin is the hormone that directs the glucose that is produced by breaking down the food into the body cells. Any change in the production of insulin leads to an increase in the blood sugar levels and this can lead to damage to the tissues and failure of the organs [3] such as kidney, eye, heart, nerves, foot etc [5]. Generally a person is considered to be suffering from diabetes, when blood sugar levels are above normal (4.4 to 6.1 mmol/L) [3].

Diabetes mellitus is classified into four broad categories: type 1, type 2, gestational diabetes and other specific types. All forms of diabetes increase the risk of long-term complications. These typically develop after many years, but may be the first symptom in those who have otherwise not received a diagnosis before that time [2]. Cause of diabetics are not yet entirely understood, scientist believe that both genetic factors and environmental triggers are involved therein [4]. Diabetes can be controlled by using different measures like insulin and diet. For this it should be identified as early as possible and subsequently provide appropriate treatment. Most of the classifying, identifying and diagnosing treatments are based on chemical and physical tests [6].

Data mining is the process of selecting, exploring and modeling large amounts of data. This process has become an increasingly pervasive activity in all areas of medical science research.Data mining has resulted in the discovery of useful hidden patterns from massive databases [7]. It is a multidisciplinary field of computer science which involves computational process [5]. Data mining is one of the "Knowledge Discovery in Databases" processes. The overall goal of the data mining process is to extract information from a dataset and transform it into an understandable structure for further use. This process has become an increasingly pervasive activity in all areas of medical science research. Data mining problems are often solved using different approaches from both computer sciences, such as multi-dimensional databases, machine learning, soft computing and data visualization; and statistics, including hypothesis testing, clustering, classification, and

regression techniques. In recent years, data mining has been used widely in the areas of science and engineering, such as bioinformatics, genetics, medicine, and education [8].

There are several major data mining techniques have been developed and used in healthcare management for, Diagnosis and Treatment, Healthcare Resource Management, Customer Relationship Management and Fraud and Anomaly Detection. Data mining techniques can help Physicians identify effective treatments and best practices, and Patients receive better and more affordable healthcare services. There are some famous data mining methods are broadly classified as: On-Line Analytical Processing (OLAP), Classification, Clustering, Association Rule Mining, Temporal Data Mining, Time Series Analysis, Spatial Mining, Web Mining etc [9].

The purpose of this study was to compare multiple prediction models for diabetes incidence based on common risk factors. This study developed three widely used data mining classification models, logistic regression, artificial neural networks (ANNs) and decision tree, along with a 10- fold cross-validation technique. Accuracy, sensitivity and specificity were used to evaluate them.

The remainder of the paper is organized as follows: In Section 2, the Literature review discussed. In Section 3, proposed structure and methodology used in this paper have been given. Section 4 deals with result analysis of data mining techniques used along with its standard tasks are presented. Further we discussed about the results and analysis of the model and conclusion.

II. Literature Review

Abdullah A. Aljumah, Mohammed GulamAhamad, Mohammad Khubeb Siddiqui (2013) discussed about predictive analysis of diabetic treatment using a regression-based data mining technique. The Oracle Data Miner (ODM) was employed as a software mining tool for predicting modes of treating diabetes. Vector machine algorithm was used for experimental analysis. Datasets of Non Communicable Diseases (NCD) risk factors in Saudi Arabia were obtained from the World Health Organization (WHO) and used for analysis. Authors conclude that drug treatment for patients in the young age group can be delayed to avoid side effects. Also, they conclude that elderly diabetes patients should be given an assessment and a treatment plan that is suited to their needs and lifestyles. In this study, predictions on the effectiveness of different treatment methods for young and old age groups were elucidated [7].

Xue-Hui Menga, Yi-Xiang Huang a, Dong-Ping Rao b, Qiu Zhang a, Qing Liu b (2013) in this study authors compare the performance of logistic regression, artificial neural networks (ANNs) and decision tree models for predicting diabetes or prediabetes using common risk factors. A standard questionnaire was administered to obtain information on demographic characteristics, family diabetes history, anthropometric measurements and lifestyle risk factors. They used three predictive models using 12 input variables and one output variable from the questionnaire information. This study may assist future researchers in choosing the optimal predictive models for implementing community lifestyle interventions to decrease the incidence of diabetes [20].

Mukeshkumari, Rajan Vohra, Anshularora (2014) inthis article predicting diabetes by applying data mining technique based on discovery of knowledge from medical datasets is important in order to make effective medical diagnosis. In this paper, Bayesian Network classifier was proposed to predict the persons whether diabetic or not. The dataset used is collected from a hospital, which collects the information of persons with and without diabetes. Pre-processing is used to improve the quality of data. The techniques of pre-processing applied are attributes identification and selection, data normalization, and numerical discretization. Next, classifier is applied to the modified dataset to construct the Bayesian model. Weka tool used to do simulation, and the accuracy of the model is calculated and compared with other algorithms efficiency [21].

J. Pradeep Kandhasamy, S. Balamurali (2015) this research study compare the performance of algorithms those are used to predict diabetes using data mining techniques. Also authors classifiers J48 Decision Tree, K-Nearest Neighbors, and Random Forest, Support Vector Machines to classify patients with diabetes mellitus. Authors compared four prediction models for predicting diabetes mellitus using 8 important attributes under two different situations. One is before pre-processing the dataset. Here the studies conclude that the decision tree J48 classifier achieves higher accuracy of 73.82 % than other three classifiers. After pre-processing, the dataset given more accurate result when compared to the previous studies. In this case, both KNN (k=1) and Random Forest performance much better than the other three classifiers and they provide 100% accuracy. From this we can come to know that after removing the noisy data from our dataset it will provide good result for our problems [2].

Sajida Perveena, Muhammad Shahbaza, Aziz Guergachib, Karim Keshavjeec (2016) Authors discussed new conventional systems, which are typically, based either just on a single classifier or a plain combination thereof. In this study follows the adaboost and bagging ensemble techniques using J48 decision tree as a base learner along with standalone data mining technique J48 to classify patients with diabetes mellitus using diabetes risk factors. Decision tree is one of the most powerful and widely applied techniques for classification and prediction which constructed reasonably good models with higher performance to classify diabetic patients, across three age groups in the Canadian population, using bagging adaboost as well as J48 decision tree. The author informed that, in future this approach can be applied on other disease datasets such as hypertension, coronary heart disease and dementia. Furthermore, diverse individual techniques like Naïve Bayes, SVM and neural networks etc. can be incorporated as base learners in ensemble framework [4].

III. Proposed architecture & Methodology

Preprocessing

Data preprocessing is a data mining technique. It is used to reduce the data. There are many data reduction techniques are available such as data compression, numerosity reduction, dimensionality reduction and discretization. In our proposed method, we have used selection attributes filters for dimension reduction to select the subset of attributes from original data to improve the classifiers accuracy.

Classification

Classification is one of the data mining Technique. It is used to group the instances which belong to same class. It is a supervised learning, in which predefined training data is available. Most popular data mining classification techniques are decision trees and neural networks. They are given below.

Neural networks

A multilayer perceptron (MLP) is a feed forward artificial neural networkmodel that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron with a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training the network[10][11][14] .

Decision tree

Decision tree is one of the classification techniques in data mining. It is tree-like graph. [14] The internal node denotes a test on attribute, each branch represents an outcome of the test, and the leaf node represent classes. It is a graphical representation of possible solutions based on condition from these solutions optimum course of action is carried out. In our work, we have used two decision tree classifier i.e., decision stump and J48 to classify the hypothyroid data set.The Algorithm of J48 and decision stump is given below.
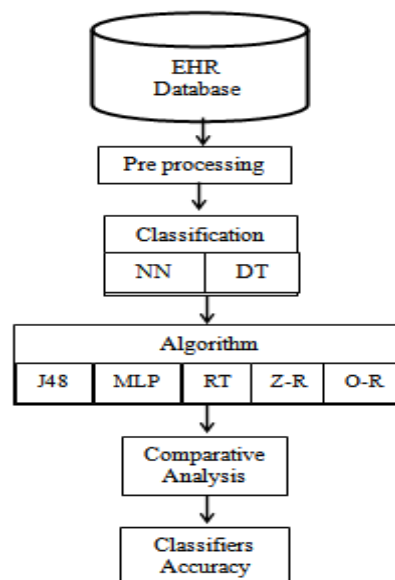

Figure1. Data processing architecture model

IV.   Algorithms

A. J48 Algorithm

J48 is a tree based learning approach. It is developed by Ross Quinlan which is based on Iterative Dichotomiser 3 (ID3) algorithm.[14] J48 uses divide-and-conquer algorithm to split a root node into a subset of two partitions till leaf node (target node) occur in tree. Given a set T of total instances the following steps are used to construct the tree structure.

Step 1: If all the instances in T belong to the same group class or T is having fewer instances, than the tree is leaf labeled with the most frequent class in T.

Step 2: If step 1 does not occur then select a test based on a single attribute with at least two or greater possible outcomes. Then consider this test as a root node of the tree with one branch of each outcome of the test, partition T into corresponding T1 , T2 , T3 ...., according to the result for each respective cases, and the same may be applied in recursive way to each sub node.

Step 3: Information gain and default gain ratio are ranked using two heuristic criteria by algorithm J48.

B. Decision Stump

A decision stump is a machine learning model consisting of a one-level decision tree. That is, it is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes. Decision stump makes a prediction based on the value of just a single input feature. Sometimes they are also called 1-rule. [14]

Random Tree

Random Tree is a supervised Classifier; it is an ensemble learning algorithm that generates lots of individual learners. It employs a bagging idea to construct a random set of data for constructing a decision tree. In standard tree every node is split using the best split among all variables. In a random forest, every node is split using the best among the subset of predicators randomly chosen at that node. Random trees have been introduced by Leo Breiman and Adele Cutler.The algorithm can deal with both classification and regression problems. Random trees is a group (ensemble) of tree predictors that is called forest. The classification mechanisms as follows: the random trees classifier gets the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of "votes". In case of a regression, the classifier reply is the average of the responses over all the trees in the forest. Random Trees are essentially the combination of two existing algorithms in Machine Learning: single model trees are merged with Random Forest ideas. Model trees are decision trees where every single leaf holds a linear model which is optimized for the local subspace explained by this leaf. Random Forests have shown to improve the performance of single decision trees considerably: tree diversity is created by two ways of randomization[13,15,20]. First the training data is sampled with replacement for each single tree like in Bagging. Secondly, when growing a tree, instead of always computing the best possible split for each node only a random subset of all attributes is considered at every node, and the best split for that subset is computed. Such trees have been for classification Random model trees for the first time combine model trees and random forests. Random trees uses this produce for split selection and thus induce reasonably balanced trees where one global setting for the ridge value works across all leaves, thus simplifying the optimization procedure. [21] [17] [18] [19]

Zero R

ZeroR is the simplest of the rule based classifiers which relies on the target and ignores all predictors[14]. It simply predicts the majority class. It is based on Frequency Table. The ZeroR classifier takes a look at the target attribute and its possible values. It constructs the frequency table and selects its most frequent value. It will ever output the value that is most frequently found for the target attribute in the given dataset. ZeroR as its names suggests; it does not include any rule that works on the non-target attributes. So more specifically it predicts the mean (for a numeric type target attribute) or the mode (for a nominal type attribute) [14].

OneR

OneR, is a short for "One Rule". It is a simple and accurate, classification algorithm that generates one rule for each predictor in the data, and then selects the rule with the smallest total error as its "one rule". To create a rule for a predictor a frequency table for each predictor against the target has been constructed. It has been shown that OneR produces rules only slightly less accurate than state-of-the-art classification algorithms while producing rules that are simple for humans to interpret [14].

V.   Experiment and Result Analysis

The open source software Waikato Environment for knowledge Analysis 3.7(WEKA) is used for experiment. It is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data preprocessing, classification, regression, clustering, association rules, feature selection and visualization. Weka can download from the website [16].

Performance Measure of Classifiers

In our experiment data is supplied to classifier of One R, Zero R, Random tree, Multi Layer Perceptron,J48 Algorithm and decision stump to classify the data. The classifiers performance is evaluated through Confusion Matrix.

Confusion Matrix

It is used for measuring the performance of classifiers. In the confusion matrix, correctly classified instances are calculated by sum of diagonal elements TP (True Positive) and TN (True Negative) and others as well as FP (false positive) and FN (False Negative) are called incorrectly classified instances.
Accuracy

It is defined as the ratio of correctly classified instances to total number of instances in the dataset.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Result Analysis

There are totally 768 records in diabetes dataset. In that 66% records are taken as training dataset. The remaining 261 records in the diabetes test dataset. All the records are classified as True positive,True negative. The following Table 2 representsconfusion matrix for decision stumpAlgorithm.

Target class	Test positive	Test Negative
Test positive	161	17
Test Negative	31	52

Table 2: Confusion matrix for decision stump

In decision stump classifier, the correctly identified instances are 213 and incorrectly identified instances are 48. The following Table 3 represents confusion matrix for J48Algorithm.

Target class	Test positive	Test Negative
Test positive	168	10
Test Negative	26	57

Table3: Confusion matrix for J48 Algorithm

In J48 classifier, the correctly identified instances are 225 and incorrectly identified instances are 36.

The following Table 4 represents confusion matrix for MLPAlgorithm.

Target class	Test positive	Test Negative
Test positive	169	9
Test Negative	26	57

Table4: Confusion matrix for MLP Algorithm

In MLP classifier, the correctly identified instances are 226 and incorrectly identified instances are 35.

The following Table 5 represents confusion matrix for One RAlgorithm.

Target class	Test positive	Test Negative
Test positive	161	17
Test Negative	31	52

Table5: Confusion matrix for One R Algorithm

In One R classifier, the correctly identified instances are 213 and incorrectly identified instances are 48.

The following Table 6 represents confusion matrix for Zero R Algorithm.

Target class	Test positive	Test Negative
Test positive	178	0
Test Negative	83	0

Table 6: Confusion matrix for Zero R Algorithm

In Zero R classifier, the correctly identified instances are 178 and incorrectly identified instances are 83.

The following Table7 represents confusion matrix for Random Tree Algorithm.

Target class	Test positive	Test Negative
Test positive	169	9
Test Negative	26	57

Table 7: Confusion matrix for Random tree Algorithm

In Random tree classifier, the correctly identified instances are 169 and incorrectly identified instances are 57.
The following table 8 shows the accuracy and time taken to build the model of classifiers before preprocessing

classifier	Accuracy	Time taken to build Model(in Seconds)
Decision stump	71.87%	0.02
MLP	75.39%	0.66
One R	71.48%	0.03
Zero R	65.10%	0
Random Tree	68.09%	0.03
J48	74.81%	0.08

Classifier	Accuracy	Time taken to build Model(in Seconds)
Decision stump	81.60%	0.02
MLP	86.59%	0.43
One R	81.60%	0.02
Zero R	68.19%	0.03
Random Tree	86.59%	0.1
J48	86.20%	0.07

Table 8: Accuracy and time taken to build model of classifiers

The following Table 9 depicts detailed accuracy and time taken to build the model of classifiers after preprocessing.
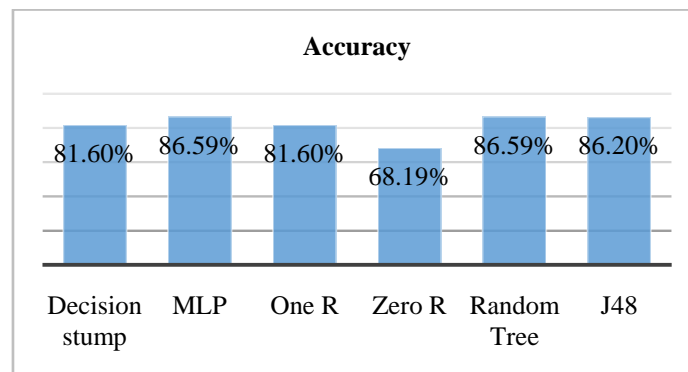

Figure 2. Accuracy of classifiers

In this chart, X axis represent the algorithm and Y axis represent the Accuracy. It shows that the Accuracy of random tree is 86.59% which is more than other classifiers.
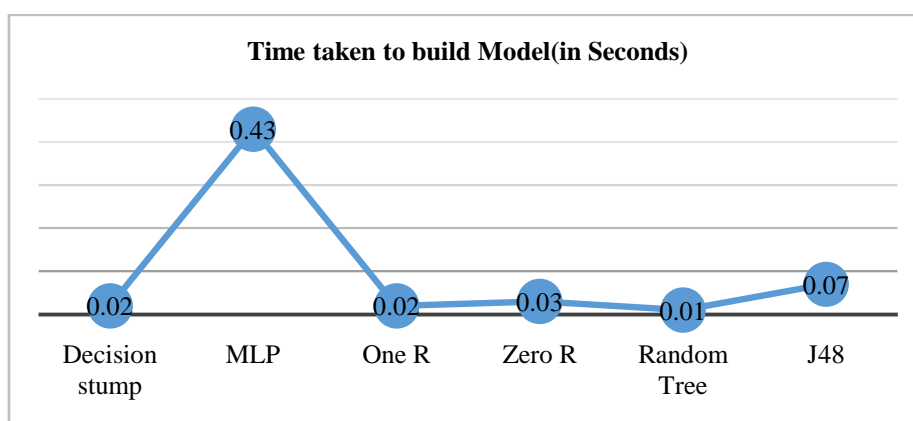The following chart 2 shows the time taken by the classifiers.

Figure 3. Time taken by the classifiers

In this chart, X axis represents the algorithm and Y axis represents the time to build the model. It shows that the time to build the model of random tree is 0.01 seconds which is less than other classifiers.

VI.   Discussion

In WEKA, there are many classification techniques available. These classification techniques are used to diagnose diabetes and some other clinical diagnosis issues. Studies showed that many researchers used different methods to diagnose the diabetes and achieved the high accuracy of classifiers for the dataset is taken from UCI machine learning repository. [Murat Koklu and YavuzUnal, 2013], proposed MLP,J48,Naivebayes classification Algorithm on diagnosing diabetes. They showed that the prediction accuracy of Naïve bayes classifier is 76.3%. . [Ashwinkumar.U.M and Dr.Anandakumar.K.R, 2012], proposed c4.5 and incremental classification Algorithm in which C4.5 gives prediction accuracy of 68%.In our study,we have used selection attribute filter for dimensionality reduction to select the subset of attributes from original data to improve the performance of classifier. We proposed five classification algorithms namely J48,One R, Zero R, Multilayer perceptron and random tree to diagnose the diabetes. The performances of classifiers are evaluated through the confusion matrix.In this experiment, the classifier random tree is givinghighaccuracy of 86.59% and minimum time to build the model of 0.01 seconds which is better than other algorithms.

VII.   Conclusion and Future Scope

Identification and diagnosis of disease is a very challenging task in the field of health care. In generally, various data mining techniques are used in decision making process. In our proposed method, we have used attribute selection filter to select the subset of attributes from original data and then we have applied J48and decision stump data mining classification techniques which are used to predict diabetes. The performances of classifiers are evaluated through the confusion matrix in terms of accuracy and execution time. The Random Tree Algorithm gives 86.59% which is providing better Accuracy than other classifier's accuracy and also Random Tree algorithmtakes very minimum time to classify data sets than other classifiers. The same procedure is used to apply for other disease datasets such as kidney disease, cervical cancer, heart disease, breast cancer, Lung cancer and so on.

References

[1].   Thirumal P. C., Nagarajan N, Utilization of data mining techniques for diagnosis of diabetes mellitus-A Case study, ARPN Journal of Engineering and Applied Sciences, Vol. 10, No. 1, January 2015.
[2].   J. Pradeep Kandhasamy, S. Balamurali, Performance Analysis of Classifier Models to Predict Diabetes Mellitus, Procedia Computer Science 47 (2015) 45 – 51, Available online at www.sciencedirect.com
[3].   AiswaryaIyer, S. Jeyalatha and RonakSumbaly, Diagnosis of diabeters using classification mining techniques, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015
[4].   Sajida Perveena, Muhammad Shahbaza, Aziz Guergachib, Karim Keshavjee, Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia Performance Analysis of Data Mining Classification Techniques to Predict Diabetes,
[5].   Dr. M. Renuka Devi, J. Maria Shyla, Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 1 (2016) pp 727-730 © Research India Publications. http://www.ripublication.com.

[6].   VeenaVijayan V. , AswathyRavikumar, International Journal of Computer Applications (0975 – 8887) Volume 95– No.17, June 2014Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus,

[7].   Abdullah A. Aljumah, Mohammed GulamAhamad, Mohammad Khubeb Siddiqui,Application of data mining: Diabetes health care in young and old patients Journal of King Saud University – Computer and Information Sciences (2013) 25, 127–136

[8].   Dr. D. Ashok Kumar, R. Govindasamy, Performance and Evaluation of Classification Data Mining Techniques in Diabetes / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 1312-1319

[9].   Rumelhart, David E., Geoffrey E. Hinton, and R. J. Williams. "Learning Internal Representations by Error Propagation". Volume 1: Foundations. MIT Press, 1986.

[10].   Jiawei Han, KamberMicheline (2009). Datamining: Concepts and Techniques, Morgan Kaufmann Publisher.

[11].   Wikipedia contributors, ―Random tree_algorithm,‖ Wikipedia, The Free Encyclopedia. Wikimedia Foundation, 28- Jan-2015.Availablefrom: http://en.wikipedia.org.[Last accessed on Nov 2].

[12].   L. Breiman, ―Random forests,‖ Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.Availablefromhttp://www.cs.waikato.ac.nz/ml/weka/.[Last accessed on Oct 24].

[13].   Breiman Leo, . "Random Forests". Machine Learning, 45 (1): 5–32, 2001.

[14].   Liaw, Andy. "Documentation for R package random forest". Retrieved 15 March 2013. U.S. trademark registration number 3185828, registered 2006/12/19.

[15].   Amit, Yali; Geman,Donald, "Shape quantization and recognition with randomized trees" (PDF). Neural Computation 9 (7): 1545–1588, 1997.

[16].   N. Landwehr, M. Hall, and E. Frank, ―Logistic model trees,‖ Mach. Learn., vol. 59, no. 12, pp. 161-205, 2005.

[17].   Murat Koklu and YauzUnal, " Analysis of a International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 1 (2016) pp 727-730 © Research India Publications. http://www.ripublication.com .

[18].   Ashwinkumar.U.M, Dr. Anandakumar K.R, "Predicting Early Detection of cardiac and Diabetes symptoms using Data mining techniques", International conference on computer Design and Engineering, vol.49, 2012.

[19].   Xue-Hui Meng a, Yi-Xiang Huang a, Dong-Ping Rao b, Qiu Zhang a, Qing Liu b, Kaohsiung Journal of Medical Sciences, Volume 29, Issue 2, Pages 93-99, , February 2013.

[20].   Mukesh kumari1, Dr. Rajan Vohra, Anshularora, Prediction of Diabetes Using Bayesian Network, International Journal of Computer Science and Information Technologies, , 5174-5178, Vol. 5 (4) , 2014.

	Dr Prakash Kuppuswamy, working atComputer Engineering & Networks Department in Jazan University, KSA. Scholar from Dravidian University, India. He published more than 30 International Research article/journals and Participated in many international Conferences at Maldives, Libya and Ethiopia, India and Saudi arabia. His research area includes Cryptography, Bio-informatics and E-commerce security, Cloud Security, Internet of Things and wireless network security etc.
	Dr Rajan John, working at**Department of** Computer Science & Engineering in Jazan University, KSA. Scholar from Bharathiyar University, India. He published more than 20 International Research article/journals and Participated in many international Conferences at Ghana and india. His area of research is Data Mining.
	Dr Shanmuasundaram marrappan, working at **Department of** Computer Science & Engineering in Jazan University, KSA. Scholar from Bharathiyar University, India. He published more than 20 International Research article/journals and Participated in many international Conferences at India and Korea. He has excellent industrial experience and specialized in Java developer. His research areas Digital Image Processing, Image Fusion, Remote Sensing and GIS, Robotics and Computer Vision