ISSN PRINT 2319 1775 Online 2320 7876

*Research paper* © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

# Analysis of dimensionality Reduction Using Singular Value Decomposition (SVD) andTwo Dimensional Haar Wavelets

**Rajesh Kumar E**

Department of CSE, Koneru Lakshmaiah Education Foundation,

Vaddeswaram, AP, India. rajthalopo@gmail.com,

**Dr. Sivakumar Selvarasu**

Department of Computer Applications, Faculty of Science and Humanities,

SRM Institute of Science and Technology, KTR Campus, Kattankulathur, Tamil Nadu - 603 203.

## Abstract

Due to advances of digital technology in all sectors ranging from healthcare, production, web organization a huge data had been generated through this domains. Machine learning algorithms are used to uncover the hidden features of these data. The main curse of these datais that, it also generate huge volume of redundant data. So we need a data reduction techniqueto reduce the data and analyze the key features hidden in the data. In this work an attempt has been made to combine Singular value Decomposition(SVD) along with Two dimensional Haar wavelet has been used to reduce the dimension of the data. To validate the above techniques we used Heart Disease Data has been used . Our experimental results further confirms that by using Two dimensional Haar wavelets along with the traditional SVD givesus better results.

## Introduction

Machine Learning (ML) is considered as one of the fast growing technology for the past two decades. It has lot of applications in various science and engineering domains. It application not only restricted to engineering it also spans various disciplines such as computer vision, bioinformatics, healthcare, banking sector, fraud detection. The main role of ML is indispensable in different areas. ML are used in different areas to predict and classify the test data to produce the accurate results. Nowadays, ML algorithms play a pivot role in health care sector, [1] Due to this a large amount of tremendous data is collected daily to update the

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

status of a certain disease. Even though large amount of data are generated through this health care , but only a portion is useful for a decision making task. Although ML can process a huge data [2] their performance will diminishes when the dimensionality of the data increases [3].

While dealing the health care data when the number of attributes increases, the number of proportionality increases and as a result it is difficult for the learned model to make the wise decisions [4]. Thus the trained model lost its significant while dealing the large number of data. Hence before dealing the huge data, one wants a dimensionality reductionalgorithm which overcome the curse dimensionality and also assures us without much loss of information. Reduction of dimensionality helps us to process a filtered data in a very effective manner.

According to the World Health Organization (WHO) Report 2015 Heart disease is considered as a leading cause of death across the globe. However the mortality rate has been decreased in high-income countries. Meanwhile the heart related death has been increased in middle-income and low income countries. Although heart disease usually affects the aged adults, the symptoms may be begin in early life. To over come the heat disease necessary steps have been taken form childhood.[5]

Heart Disease is one of the diseases that affects the function of heart, blood vessels or both. The most common cause of the disease are atherosclerosis and/or hypertension. Athericids is a condition that develops when a substance called plaque builds up in the wall of the arteries. This buildup narrows the arteries, making it harder for blood to flow through. If a blood clot appears, it can stop the flow of blood. This can eventually results in heart attack or stroke. [6, 7] There are several major factor for heart disease such as age, gender, diabetes mellitus, tobacco smoking, diabetes excessive meat consumption, excessive alcohol consumption, family history, air pollution, lack of physical activity, psychosocial factors. Etc.,

While dealing the clinical data the patient may provide data which may not be irrelevant to the prediction of the heart disease. It eventually ends up with irredundant data. So one wants an efficient tool to deal with the feature extraction and gives a better result for the prediction

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

of the heart disease. In literature [8] there are several dimensionality reduction that is available for the feature extraction of the available data such as PCA, KPCA, LDA , LLE and SVD.

In this paper an attempt has been made on "Heart Disease Data Set" obtained from the UCI Machine Repository.[9] First SVD is applied on the raw data for data reduction later after obtained the confusion matrix two dimensional Discrete Haar wavelet are obtained.

**Dimensionality Reduction**

Dimensionality reduction is a one of the techniques which extract the important features thereby reducing the size of the dataset without affecting the characteristics of data. IN [10] have used PCA and CCA for analysis of heart disease. This technique provides better data visualization, data compression, improved classification accuracy, fast and efficient data retrieval. The main merits of the dimensionality reduction technique are reducing the dimension enhances the prediction accuracy of the classifier with good performance and also reduces the computational cost which was shown in Fig. 1

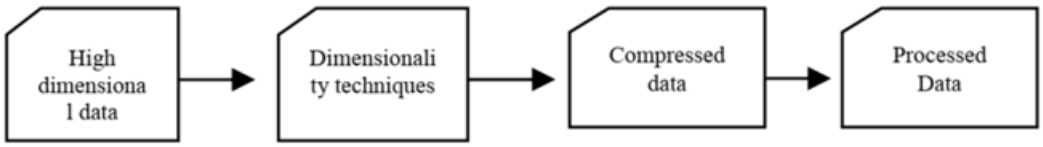



Fig. 1. Dimensionality Reduction Process

In this work some classifiers in ML like decision Tree, Naïve Baye, Random Forest and SVM are used to classify the data.

**Singular Value Decomposition (SVM)**

Singular Value Decomposition provides an exact representation of a dataset represented by a matrix with any number of dimensions. If we select the less number of dimensions we may arrive a more precise on the data set. The SVD will select "m" largest singular value which is chosen according the theorem [11,12,13]

$$Y \rightarrow U.M.V^{T} \quad (1)$$

Here

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

- Y , represents the original matrix, which is decomposed into three matrices.
- U is an (n x k ) matrix with perpendicular unit columns such that $U.U^T = U^TU = 1$.
- M is square matrix of order (k x k) diagonal matrix, where the entries of the main diagonal contains the singular values and the rest of the elements are zero.
- V is an orthogonal matrix ($V^T.V = V.V^T = I$) of order ( k x d), known as singular values.

The main aspect of the SVD Theorem rebuild the given input matrix Y in a k-lowdimensional / rank and the singular value of V are positive and arranged in descending order. SVD is also considered as one of the preferred dimensionality reduction, for prediction problems [14, 15,16]

**Algorithm Input** $Y \in R^{n \times d}$

**Output** $Z \in R^{n \times k}$

1. Decompose Y into 3 matrices; as U, S and V
2. Build Z be selecting k top singular values form V

**Two Dimensional Wavelets**

Wavelets are considered one of the best tool for signal processing, image processing . There are several wavelet are available in the literature like Haar wavelet, Daubechies wavelets, Coifman wavelet, Meyer wavelets.[17,18,19,]. Depends upon the applications of the problem different wavelets are used. One of the simplest wavelet is the Haar wavelet which was derived from the Haar scaling function and defined as

$$\varphi(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & otherwise \end{cases}$$

Similarly one can define the two dimensional wavelets whose scaling function and vertical, horizontal, and diagonal scaling function is defined as follows

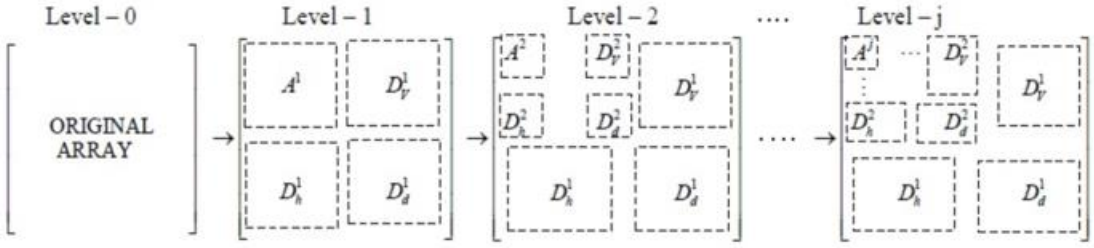
ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

Using the above coefficients we can apply two dimensional discrete Haar wavelet transform. When the consider the data which is of the following matrix can be represent by

$$S = \begin{bmatrix} s_{11} & s_{12} & s_{13} & s_{14} & \cdots & s_{1(n-1)} & s_{1n} \\ s_{21} & s_{22} & s_{23} & s_{24} & \cdots & s_{2(n-1)} & s_{2n} \\ s_{31} & s_{32} & s_{33} & s_{34} & \cdots & s_{3(n-1)} & s_{3n} \\ s_{41} & s_{42} & s_{43} & s_{44} & \cdots & s_{4(n-1)} & s_{4n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ s_{(n-1)1} & s_{(n-1)2} & s_{(n-1)3} & s_{(n-1)4} & \cdots & s_{(n-1)(n-1)} & s_{(n-1)n} \\ s_{n1} & s_{n2} & s_{n3} & s_{n4} & \cdots & s_{n(n-1)} & s_{nn} \end{bmatrix}$$

Where S is the square matrix of order n x n . where ($n = 2^j$ ) j represent the level of decomposition of the matrix. The input matrix in this case denotes the feature vector can be decomposed by applying wavelet coefficient as into a submatrix as shown in the figure.




Decomposing of original matrix into various levels.

The $A^1$, $D^1$, $D^1_v$, $D^1_h$ and $D^1_d$ are obtained by applying scaling approximating , vertical wavelet coefficient , horizontal wavelet coefficient, and diagonal wavelet coefficient.

**Data description**

"Heart Disease Data" obtained from UCI Machine Learning Repository consists of 4 data bases. All Attributes are numeric – valued. The data was collected from four location Cleveland Clinic Foundation, Hungarian Institute of cardiology, V.A Medical centre and University hospital giving four datasets. Each data set consist of identical features. Clinical symptoms of heart disease represented by 10 variables are used for analysis. [20 -22] All 4 datasets are combined which gives a total 920 samples, with 411 healthy and 509 unhealthy samples. Removal of data with missing values reduced to a total of 686 samples with 352 samples of healthy and remaining 334 of unhealthy patients.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

| S.No | Attribute | Description | Values |
|---|---|---|---|
| 1 | Age | In Years | Range 28 to 75 |
| 2 | Gender | Male/Female | 1/0 |
| 3 | CP – Chest pain Type | Typical / Atypical /Non -angina / Asymptotic | 1/2/3/4 |
| 4 | TRESTBPS | Resting Bp (mmHg) | Range 0 to 200 |
| 5 | CHOL | Choleterol (mg /dI) | Range 0 to 603 |
| 6 | FBS | Fasting Blood Sugar >120 mg/dl F/T | 0 /1 |
| 7 | RESTECG | Resting ECG results Normal / ST abnormal / left ventricle hypertrophy | 0/1/2 |
| 8 | THALACH | Max Heart rate achieved | Range 69 to 202 |
| 9 | EXANG | Exercise induced angina No /Yes | 0/1 |
| 10 | OLDPEAK | ST depression induced by exercise | 0/1 |
| 11 | CLASS | Healty / Unhealthy | 0/1 |

## PROPOSED METHODOLOGY

This paper explores the effect of feature extraction and dimensionality reduction techniques on the performance of ML algorithms on Heart disease data set. The variousstep used in this work are discussed below.

Step1:

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

Feature extraction technique, normalization and conversion of categorical data to numerical data is applied on Heart disease data set. To normalize the input dataset ,standard scalar normalization method is used.Step 2:

The normalised dataset is tested using ML algorithms, Decision Tree, Naïve Bayes, Random Forest and Support Vector Machine (SVM). The performance of these classifiersis then evaluated on the metrics, Precision, Accuracy, Sensitivity and Specificity

Step 3:

SVD is applied on the normalized dataset to extract the most important features. The resultant dataset is then tested using the ML algorithms.

Step 4.

After obtained the confusion matrix two dimensional discrete Haar wavelets are obtainedfor calculate the performance of the various measures are evaluated.

**Metrics for Evaluation of the Model**

It is the percentage of correct predictions that a classifier has made when comparted tothe actual value of the label in testing phase.

The metric used in the analysis are Accuracy, Sensitivity, and SpecificityAccuracy can be calculated using the following formula

Accuracy = (TN +TP)/(TN + TP + FN +FP) where, TP is true positives, TN representstrue negatives, FP is false positives, FN is false negatives

Sensitivity

It is the percentage of true positives which are truly identified by the classifier duringtesting. It can be derived using the give formula TP/(TP + FN)

Specificity

It can calculate the percentage of true negatives that are correctly identified by the classifier during testing and is derived using the following TN/(TN + FP)

**Results and Discussions**

The test is performed on Heart disease dataset which is available UCI Machine learning

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

repository using Python 3. A laptop with windows 10 operating system and 8GB RAM is used for this experiment.

Initially some classifiers like Decision Tree, Naïve Bayes, Random forest and SVM are used and the confusion matrix are listed below. The confusion matrices shows that the SVM and Random forest algorithms performs better than Decision tree and Naïve Bayein terms of precision.

Decision Tree Confusion Matrix: [3 53 0 ]
0 0 41

Decision Tree Confusion matrix. :Table 1

| | Precision | Recall | F-1Score |
|---|---|---|---|
| 1.0 | 0.95 | 0.95 | 0.95 |
| 2.0 | 0.99 | 0.93 | 0.94 |
| 3.0 | 1.00 | 1.00 | 1.00 |

Naïve Bayes Confusion Matrix: [330 2 3
3 53 0 ]
0 6 41

Naïve bayes Confusion matrix. Table 2

| Precision | Recall | F-1Score |
|---|---|---|
| 0.97 | 0.97 | 0.97 |
| 0.89 | 0.93 | 0.90 |
| 0.92 | 0.91 | 1.00 |

Random forest Confusion matrix. Table 3

| Precision | Recall | F-1Score |
|---|---|---|
| 0.98 | 0.97 | 0.97 |

ISSN PRINT 2319 1775 Online 2320 7876

Research paper

© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

| | | |
|---|---|---|
| 0.89 | 0.93 | 0.90 |
| 0.97 | 0.91 | 0.90 |

SVM Confusion Matrix: [3 55 0 ]

0 6 40

SVM Confusion matrix. Table 4

| Precision | Recall | F-1Score |
|---|---|---|
| 0.95 | 0.96 | 0.97 |
| 0.85 | 0.96 | 0.90 |
| 0.91 | 0.91 | 0.99 |

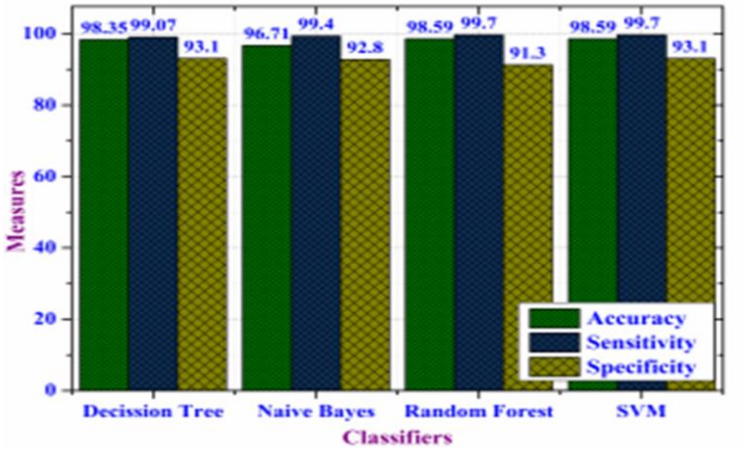



Figure 2

The above figure 2 depicts performance of the classifiers and it concludes that Decision tree and SVM perform better that the other two algorithms in terms of specificity.

Later the dimensions of the dataset is applied using SVD. The number of attributes 13 are reduced to six features and the top six features are once again subjected to two dimensional discrete Haar wavelets and the confusion matrices for these experiments are shown below

Decision Tree-SVD-2DDHW Confusion Matrix: [ 324 2 0

8 49 1 ]

0 0 42

ISSN PRINT 2319 1775 Online 2320 7876

*Research paper*

© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

Table 5: SVD -2DDHW Confusion Matrix

| Precision | Recall | F-1Score |
|---|---|---|
| 0.98 | 0.99 | 0.98 |
| 0.96 | 0.84 | 0.90 |
| 0.98 | 1.00 | 0.99 |

Naïve Bayes-SVD-2DDHW Confusion Matrix: [ 326 0 0
39 1]
19
0 42 0

Table 6 ; Naïve Bayes-SVD-2DDHW Confusion Matrix

| Precision | Recall | F-1Score |
|---|---|---|
| 0.984 | 1.00 | 0.97 |
| 0.48 | 0.67 | 056 |
| 0.97 | 1.00 | 0.99 |

Random forest-SVD-2DDHW Confusion Matrix: [ 326 0 0
39 1]
19
0 42 0

Table 7:Random Forest-SVD-2DDHW Confusion Matrix

| Precision | Recall | F-1Score |
|---|---|---|
| 0.98 | 0.99 | 0.97 |
| 0.96 | 0.84 | 0.90 |
| 0.98 | 1.00 | 0.99 |

ISSN PRINT 2319 1775 Online 2320 7876

*Research paper*

© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

Table 8 SVM-SVD-2DDHW Confusion Matrix:

| Precision | Recall | F-1Score |
|---|---|---|
| 0.98 | 1.00 | 0.98 |
| 0.96 | 0.84 | 0.90 |
| 1.00 | 1.00 | 1.00 |

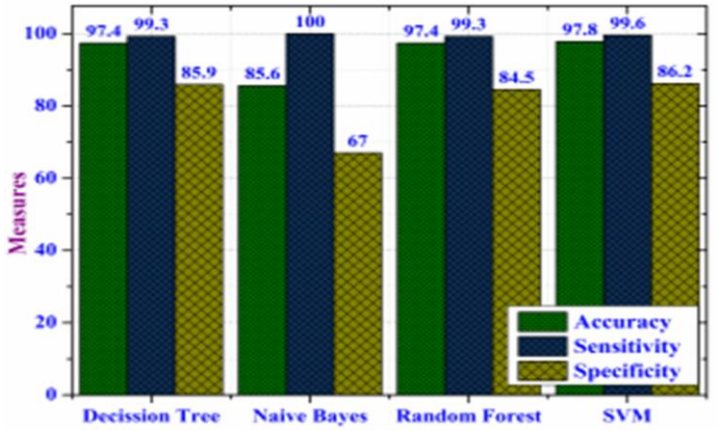



**Figure 3 Classifiers using SVD and 2DDHW**

The above figure 3 depicts the performance of these classifiers on the diminished dataset in terms of accuracy, sensitivity and specificity measures. The Accuracy of Decision tree, Naïve Bayes, Random forest and SVM are 97.4%, 85.6%, 97.4% and 97.8% respectively. Similarly the sensitivity for the above classifiers are 99.3%,100%, 99.3% and 99.6% . Finally Specificity is achieved as 85.9%,67% 84.5% ,86.2% for these classifiers. These results shows that Naïve Baye with SVD-2DDHW performs relatively less in terms of accuracy and sensitivity, whereas sensitivity is 100%.

## Conclusions

In this paper, we proposed the use of dimensionality reduction techniques with machine learning classifiers to predict whether a patient was heart disease or not. The results presented by the SVD were marvelous. The features that were present in all the data sets are Chest pain (cp), cholestrol (chol) and ST depression. SVD along with 2DDHW offers better results for prediction of chest pain for the Heart diseases data set.

ISSN PRINT 2319 1775 Online 2320 7876

*Research paper* © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

**References:**


1. C.M. Bishop, "Pattern Recoginition and Machine Learning, New York, NY, USA,Springer , 2006,
2. C.T.Rasmussen, "Gaussian Processing in Machine Learning", in Summer School onMachine Learing, Berlin, Germany, Springer, 2003, pp 63 -71
3. L. Van Der Matin, J. Van Den Herik, "Dimensionality Reduction : A comparative",J.mach. Learn Res. Vol 10 nos. 66 -71.
4. A.Zheng and A.Casari "Feature Engineering for machine learning, Principles and Techniques for Data Scientists", Newton, MA, USA, O Reilly Media, 2018,
5. S.Sivakumar, P.Ganesan, and S.Sundar "An MMDBM Classifier with CPU and CUDA GPU Computing in various sorting procedures", The International Arab Journal of Information Technology, 2017, 897 -906.
6. S Sivakumar, S Vidyanandini, E Rajesh Kumar, D Haritha, CMAK ZeelanBasha "A New And Fast Supervised Learning Algorithm Based On Blood Pressure (Bp) Data Analysis." Ilkogretim Online 2000
7. E Sreedevi, V PremaLatha, Y Prasanth, S Sivakumar "A novel Esemble Learning for Defect detection method for uncertain data" 67-79 , 2002, IGI Global.
8. Farzana Anowar, Samira Sadoui, Bassant Selim, "Conceptual and emperical comparison of dimensionality reduction algorithms", Computer Science Review 40pp 1 -12,
9. UCI Machine Learning Repository: Heart Disease Data Set online available www.achiveve.ics.uci/ml/dataset/heart +disease.
10. Tarun Kumar Gupta, Chanchal Kumar, Shiv Prakash and Mukesh Prasad: Dimensionality Reduction Techniques and its Applications: Computer Science Systems Biology, (2015)
11. H.Abdi, Singular Value decomposition and Genaralised Singular Value decomposition, I Encyclopedia of Measurement and Statistic, 2007,pp 907 -912.
12. A.G.Akritas, G.I.Malaschonok, Applications of Singular – Value Decomposition

ISSN PRINT 2319 1775 Online 2320 7876

*Research paper* © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

(SVD), Math. Comput.Simul. 67, 2004, 15 – 31.

13. J.Leskovec,A.Rajaraman, J.Ullman, Dimensionlality Reduction , in Mining of Masive Datasets, 2014, pp. 415 – 447.
14. T.Georgeena.S. Thomas, Siddhesh.S. Budhkar, Siddhesh.K.Cheulkar, .B.Choudhary, Rohan Singh:" Heart Disease Diagnosis System Using Apriori Algorithm:" International Journal of Advanced Research in Computer Science and Software Engineering,Volume 5, Issue 2, February (2015).
15. Swati A sonawale & Roshani Ade: Dimensionality Reduction:An Effective Technique for Feature Selection : International Journalof Computer Applications, Volume 117-No 3, May (2015).
16. R. Chitra1 and V. Seenivasagam : Review Of Heart DiseasePrediction System Using Data Mining And Hybrid IntelligentTechniques: Ictact Journal On Soft Computing, Volume: 03, Issue: 04, July (2013)
17. G.Hariaharan, Wavelet Analysis- An overview, Wavelet Solutions for Reaction–Diffusion Problems in Science and Engineering, 2000, 15 – 31.
18. V.Sumathy, S.Hemalatha, B.Sripathy, Comparitive Study of Two dimensional Legendre and Chebyhev Extended case, 2019, 13 – 17
19. P Vijayaraju, B Sripathy, D Arivudainambi, S Balaji "Hybrid memetic algorithm with two-dimensional discrete Haar wavelet transform for optimal sensor placement", 2017 2267 – 2278.
20. Beant Kaur, Williamjeet Singh: Review on Heart Disease Prediction System usingData Mining Techniques: International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 2 Issue: 10, October (2014)..