International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

# Human Heart Disease Prediction System Using Random Forest Technique

**H. Kaur[1*], D. Gupta[2]**

[1] CSE, I.K.G, Punjab Technical University, Kapurthala, Punjab, India
[2]CSE, I.K.G, Punjab Technical University, Kapurthala, Punjab, India

**Corresponding Author: heavenpreetkaur@gmail.com, Tel.: +91-9501098555*

**Available online at: www.ijcseonline.org**

Accepted: 16/Jul/2018, Published: 31/Jul/2018

***Abstract*—** Data mining is the analytical process to explore specific data from large volume of data. It is a process that finds previously unknown patterns and trends in databases. This information can be further used to build predictive models. The main objective of our paper is to learn data mining techniques which can be used in the prediction of heart diseases using any data mining tool. Heart is the most vital part of the human body as human life depends upon efficient working of heart. A Heart disease is caused due to narrowing or blockage of coronary arteries. This is caused by the deposition of fat on the inner walls of the arteries and also due to build up cholesterol. Thus, a beneficial way to predict heart diseases in health care industry is an effective and efficient heart disease prediction system. This system will find human interpretable patterns and will determine trends in patient records to improve health care. In this paper, Random Forest technique is applied to enhance the accuracy of the system.

***Keywords*—** Data Mining Technique, KNN, Random Forest, Heart Diseases**.**

## I. INTRODUCTION

Data mining is the process of analyzing large sets of data and extracting the meaning/pattern in data. It helps in predicting future trends and patterns, allowing business in decision making. Data mining applications can answer business questions that take much time to resolve traditionally. Large amount of data which is generated for the prediction of heart disease is analyzed traditionally and is too complicated and voluminous to be processed.

Data mining is the process of analyzing data from different views and summarizing it into useful data. "Data mining, also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories or data streams.

### A. *Data Mining Process*

Data mining is also known as Knowledge Discovery in Database, refers to finding or "mining" knowledge from large amounts of data. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. So, many people use the term "knowledge discovery in data" or KDD for data mining.

In Data mining, Knowledge extraction or discovery is done in sequential steps as in Fig 1.

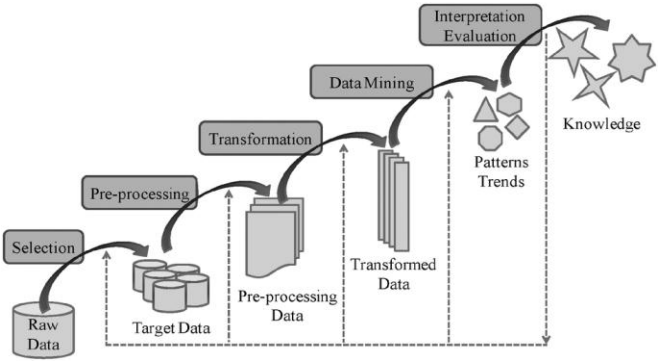


Figure 1 Data Mining Process

i) Data cleaning: This is the first step to eliminate noise data.
ii) Data integration: Data sources are combined into meaningful and useful database.
iii) Data Selection: In this data relevant to the analysis are retrieved from other various resources.
iv) Data transformation: In this data is converted or consolidated into required forms such as smoothing, normalization or aggregation.

 

v) Data Mining: In this step, various techniques and tools are applied in order to extract data pattern or rules.
vi) Pattern evaluation: In this, Attractive patterns representing knowledge are identified based on given measures.
vii) Knowledge representation: This is the last stage in which, visualization and knowledge representation technique.

B. *Data Mining Techniques*

Data mining techniques have been developing and using in data mining projects. Data mining process is taking out the information from large data sets and transforms it into some understandable form.

*Predictive Model*

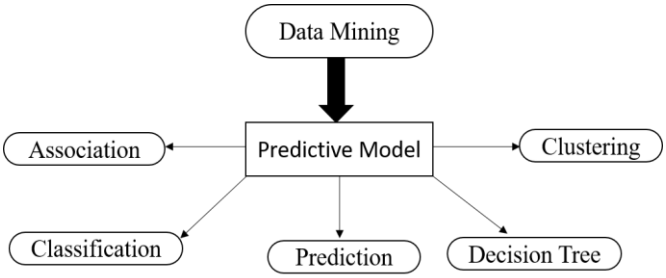


Figure 2. Data Mining Techniques

*Association*

The association technique is used to extract the relationships between attributes and items. It is common in establishing a form of statistical relationships among different interdependent variables of data mining; association rules are useful for analyzing and predicting customer behaviour.

*Classification*

Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups.It is the method that makes use of mathematical techniques such as decision trees, linear programming, neural network, and statistics. we develop the software that can learn how to classify the data items into groups.

*Prediction*

Prediction involves analyzing trends, classification, pattern matching and relation.It is one of the data mining techniques that discovers the relationship between independent variables and relationship between dependent and independent variables.

*Decision Tree*

A Decision tree is one of the most commonly used data mining techniques because its model can be easily understood by users. The decision tree simple to understand and interpret. Allow the addition of new possible scenarios. It is help to determine worst, best and expected values for different scenarios.

*Clustering*

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

Data mining provides the techniques and methods for the transformation of data into useful information for decision making. These techniques make the process fast and it takes less time for the prediction system to predict heart disease with more accuracy. In proposed work we surveyed different papers in which one or more algorithms of data mining used for the prediction of heart disease [1]. Data mining is the analytical process to explore specific data from large volume of data. It is a process that finds previously unknown patterns and trends in databases. This information is further used to build predictive models. The main objective of our paper is to learn the different data mining techniques which are used in the prediction of heart diseases using any data mining tool. Heart is the most vital part of the human body as life is dependent on efficient working of heart. A Heart disease is caused due to narrowing or blockage of coronary arteries. This is caused by the deposition of fat on the inner walls of the arteries and also due to build up cholesterol [2].

## II. RELATED WORK

**Jesmin** et al. [3] in the year 2013 in heart prediction system used Naïve Bayes and achieved 92.08 percent of accuracy. Then again he used SMO, AdaBoostM1, J48 and PART achieving 94.04%.

**M. Ambarasi** et al. [4] in the year 1999 used classification via clustering naïve bayes and achieved 88.3 percent of accuracy. Then he used decision tree used in the prediction of heart disease prediction system.

**Matjaz** et al. [5] in the heart prediction system used exercise ECG AND myocardial scintigraphy i.e. neural network and achieved 85 percent of accuracy in the year 1999.

 

**N. Aditya Sundar** et al. [6] in the year 2012 used Naïve Bayes in heart prediction system .Then he applied WAC technique..

**T.John** et al. [7] in the year 2012 used multilayer algorithm in heart prediction system and achieved 78.88 percent of accuracy. Then he applied Naïve Bayes and achieved 85.18 percent.

**Carlos** et al. [8] in the year 2001 used only one technique that was association rule in heart prediction system to diagnose heart diseases using 25 attributes in the heart data set.

## III. METHODOLOGY

Presently various algorithms are available for clustering the proposed data, in the existing work they used K nearest neighbor algorithm for Heart disease prediction system and achieved the accuracy of 73%. As we can see that there is vast scope of improvement in our proposed system, Several Parameters has been proposed for heart disease prediction system but there have been always a need for better parameters or algorithms to improve the performance of heart disease prediction system. To improve and enhance the accuracy of the system for use Random forest technique **Main steps follow in this methodology are as follow**:

**1. Initialize the dataset:** The dataset is mined, uploaded and transformed into the required matrix form with the help of data mining tool Matlab.

**2. Data Preprocessing:** the dataset contains quotes and we have used removed quote function to process the data. Further data is converted into nominal form.

**3 Apply k nearest neighbor algorithm and evaluate accuracy:** K nearest neighbor algorithm is applied to the dataset and accuracy is calculated. This algorithm assigns the object to a class which is most common in its neighbors.

**4. Apply Random Forest algorithm and evaluate accuracy:** Random Forest algorithm is applied to the dataset and accuracy is calculated. This Algorithm collect the bunches of trees and divided into parts.

**5. Prediction:** To classify the data, random forest method is used. Data is partitioned using cross validation function and different decision tree are used to learn the target variable and predict function is used to give real time prediction. Here accuracy of the system is also calculated.

**Proposed Methodology is done in sequential steps as in Fig.3**

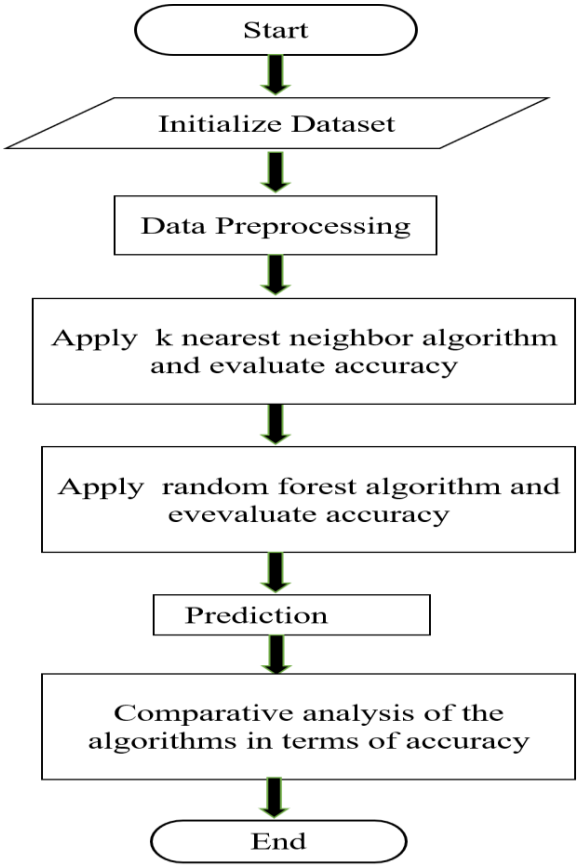


Figure .3 Flowchart of Proposed Methodology

**6. Comparative analysis of the algorithms in terms of accuracy:** Comparative analysis of the entire algorithms is done and the result of performance is calculated in terms of accuracy.

### A. KNN Classification Algorithm

In pattern recognition field, KNN is one of the most important non-parameter algorithms [11] and it is a supervised learning algorithm. The classification rules are generated by the training samples themselves without any additional data. The KNN classification algorithm predicts the test sample's category according to the K training samples which are the nearest neighbors to the test sample, and judge it to that category which has the largest category probability.[14]

- The process of KNN algorithm to classify sample X is [14]: Suppose there are j training categories $C_1$,$C_2$,…,Cj and the sum of the training samples is N

 

after feature reduction, they become m-dimension feature vector.

- Make sample X to be the same feature vector of the form (X1, X2,…, Xm), as all training samples.
- Calculate the similarities between all training samples and X. Taking the $i^{th}$ sample di ($d_{i1}$,$d_{i2}$,…,$d_{im}$) as an example, the similarity SIM(X, $d_i$) is as following:

$$sim(x,Di)=\frac{\sum_{j=1}^{m} xj.dij}{\sqrt{(\sum xj)2}\sqrt{(\sum dij)2}}$$ [14]

- Choose k samples which are larger from N similarities of SIM(X, di), (i=1, 2,…, N), and treat them as a KNN collection of X. Then, calculate the probability of X belong to each category respectively with the following formula.

$$P(x,cj)=\sum Sim\,(x,d).y(di,cj)$$ [14]

Where y(di, Cj) is a category attribute function, which satisfied.

$$y(d,cj)=\begin{cases}1, dj \in cj\\0, dj \in cj\end{cases}$$ [14]

Judge sample X to be the category which has the largest P(X, Cj).

**Flowchart of KNN Methodology is done in sequential steps as in Fig.4**

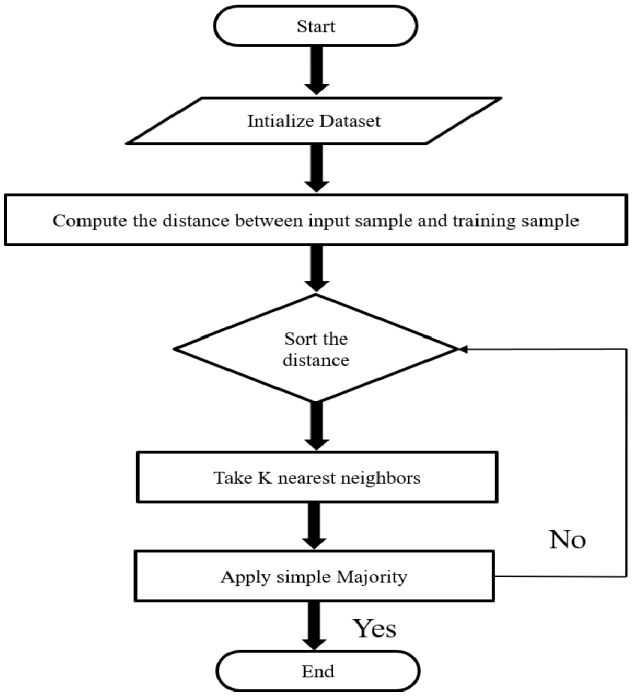


Figure.4 Flowchart of KNN Methodology

**B. Random Forest Classification Algorithm**

Each tree is constructed using the following algorithm:

1. Let the number of training cases be N, and the number of variables in the classifier be M.[20]

2. We are told the number m of input variables to be used to determine the decision at a node of the tree; m should be much less than M.[20]

3. Choose a training set for this tree by choosing N times with replacement from all N available training cases (i.e. take a bootstrap sample). Use the rest of the cases to estimate the error of the tree, by predicting their classes.[20]

4. For each node in the tree, randomly choose m variables on which to base the decision at that node. Calculate the best split based on these m variables in the training set.[20]

5. Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier).[20]

**Flowchart of Random Forest technique Methodology is done in sequential steps as in Fig.5**

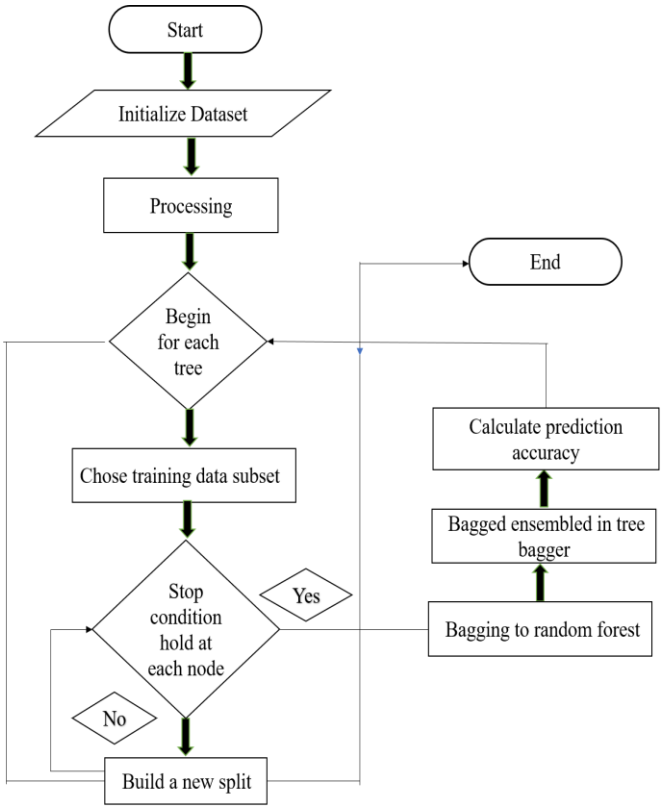


Figure.5 Flowchart of Random Forest Methodology

 

A. Decision Tree Learning

Decision trees are a popular method for various machine learning tasks. Tree learning comes from closest to meeting the requirements for serving as an off-the-shelf procedure for data mining.[17]

B. Tree bagging

Main Brief of Tree bagging is Bootstrap aggregating.

Bootstrap aggregating, also called bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting.[18]The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set $X = x_1, ..., x_n$ with responses $Y = y_1, ..., y_n$, bagging repeatedly ($B$ times) selects a random sample with replacement of the training set and fits trees to these samples:[18]

For $b = 1, ..., B$:

1. Sample, with replacement, $n$ training examples from $X$, $Y$; call these $X_b$, $Y_b$.[18]
2. Train a classification or regression tree $f_b$ on $X_b$, $Y_b$.[18]

After training, predictions for unseen samples $x'$ can be made by averaging the predictions from all the individual regression trees on $x'$:[18]

$$f = \frac{1}{B}\sum_{b=1}^{B} Fb(X') \quad [18]$$

Taking the majority vote in the case of classification trees.[18] **a. List of Heart Disease Attributes:**

| Id. | Attribute |
|---|---|
| 1. | Age |
| 2. | Blood cholesterol |
| 3. | Blood pressure |
| 4. | Hereditary |
| 5. | Smoking |
| 6. | Alcohol intake |
| 7. | Physical activity |
| 8. | Diabetes |
| 9. | Diet |
| 10. | Obesity |
| 11. | Stress |
| 12. | Gender |
| 13. | Drinker |

*Table 1- heart disease attributes*

## IV. RESULTS AND DISCUSSION

In this work, MATLAB is used to evaluate the results. In the existing work, KNN was used which has accuracy of 73.33% as shown in Fig. 3.

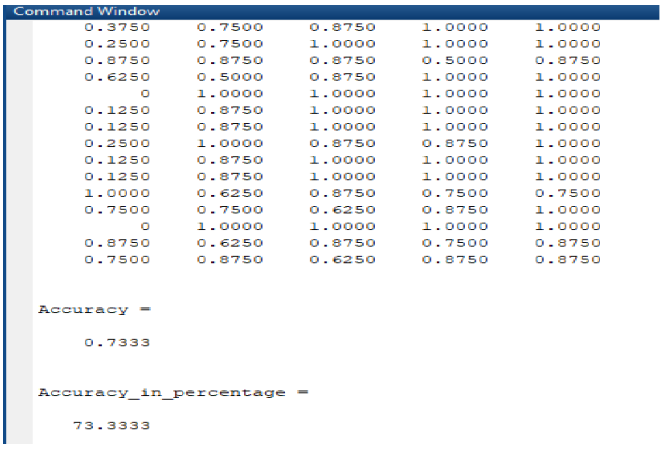


**Figure.6 Accuracy calculation of KNN**

In Figure.6 depicts that Real time prediction is done using KNN. Predict which provides predicted label which are compared with testing label**.**

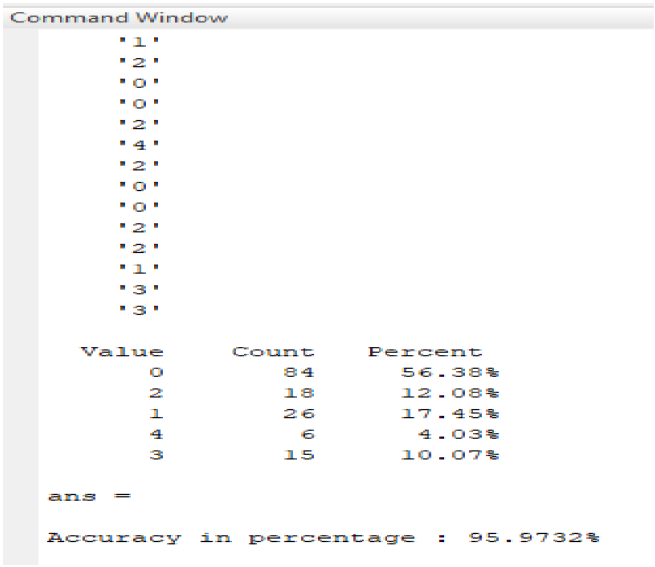


**Figure. 7 accuracy calculation of Random Forest**

Figure. 7 depicts that Real time prediction is done using Random Forest. Predict which provides predicted label which are compared with testing label**.**

 

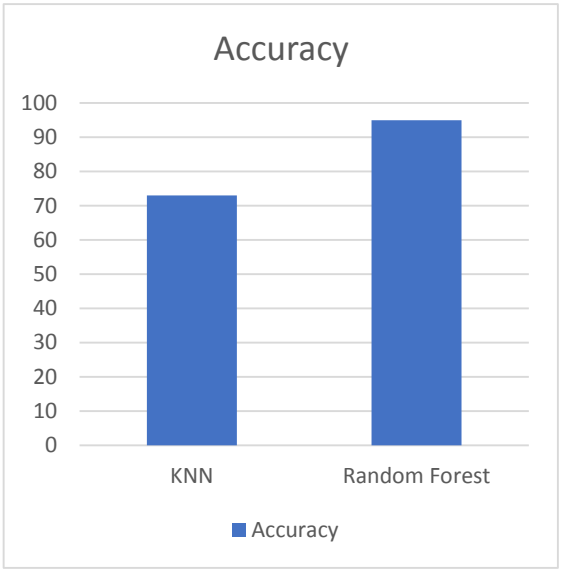


Figure.8 Compared Accuracy

Figure.8 depicts that accuracy of Random Forest is 95% as compared to KNN. From the graph it can be concluded that prediction done by random forest is much better than KNN. Because the learning rate of Random Forest is very high and is called quick learner whereas KNN is referred as lazy learner

## V. CONCLUSION AND FUTURE SCOPE

Medical related information is highly voluminous in nature in the healthcare industry. It can be derived or retrieved from various sources which are not entirely applicable in this feature. In this work, heart disease prediction system was developed using classification algorithms through Matlab data mining tool to predict effective and accurate results regarding whether the patient is suffering from heart disease or not. In future work, we have planned to propose more effective heart disease prediction system to predict heart diseases with better accuracy using different data mining techniques and compare the performance of algorithm with other related data mining algorithms.

## REFERENCES

[1] Bhupendra Kumar Jain , Manish Tiwari, "Prediction Analysis Technique based on Clustering and Classification", International Journal of Computer Sciences and Engineering, . Vol.6 , Issue.6 , pp.688-692, Jun-2018.

[2] Era Singh Kajal Ms. Nishika "Prediction of Heart Disease using Data Mining Techniques" Volume2, Issue3, pp-1-5 2012.

[3] Indira S. Fal Dessai" Intelligent Heart Disease Prediction System Using Probabilistic Neural Network" International Journal on Advanced Computer Theory and Engineering, VOL.8 No.8, August 2008.

[4] Bahadur Patel, Ashish Kumar Sen D P, Shamsher Shukla, "A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level", International Journal Of Engineering And Computer Science ISSN: 2319-7242, Page No. 2663-2671, Volume 2 Issue 9 Sept., 2013.

[5] Giorgio Barbareschi and Robbert Sanderman et al, "Socioeconomic Status and the Course of Quality of Life in Older Patients with Coronary Heart Disease", International Journal of behavioral Medicine, Vol.16, PP.197-204, 2009.

[6] Jesmin Nahar and Tasadduq Imam et al," Association rule mining to detect factors which contribute to heart disease in males and females", Journal of Expert Systems with Applications Vol.40, PP.1086–1093, 2013

[7] Thomas, J., and R. Theresa Princy. "Human heart disease prediction system using data mining techniques." 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT]

[8] Daisy JA Janssen and Emiel FM Wouters et al, "Self-perceived symptoms and care needs of patients with severe to very severe chronic obstructive pulmonary disease, congestive heart failure or chronic renal failure and its consequences for their closest relatives: the research protocol", Journal of BMC palliative care, Vol.7, 2008.

[9] Abhishek Taneja, "Heart Disease Prediction System Using Data Mining Techniques" Oriental Scientific Publishing Co., India, Vol.11, 2013.

[10] Srinivas K, Raghavendra Rao R, Govardhan A, "Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques", The 5th International Conference on Computer Science & Education Hefei, China. Volume 2, Issue 10, October 2012.

[11] Rajalaxmi R R, "A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier", International Conference on Recent Trends in Computational Methods, Communication and Controls (ICON3C 2012) International Journal of Computer Applications (IJCA), 2012.

[12] Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012.

[13] Chitra R and Seenivasagam V, "Review of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Technique", ISSN: 2229-6956(online) ICTACT Journal On Soft Computing, JULY 2013, VOLUME: 03, ISSUE: 04, 2013

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

[15] N. Aditya Sundar, P. Pushpa Latha and M. Rama Chandra, "Performance analysis of classification data mining techniques over heart disease data base", International journal of engineering science & advanced technology, Volume-2, Issue-3, May-June 2012, pp. 470 – 478.

[16] T. John Peter, K. Somasundaram, study and development of novel feature selection framwork for Heart disease prediction, International Journal of Scientific and Research Publications, April 2012, Vol. 8 .

[17]Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). The Elements of Statistical Learning (2nd ed.). Springer. ISBN 0-387-95284-5.

[18]Shinde, Amit, Anshuman Sahu, Daniel Apley, and Ge.orge Runger. "Preimages for Variation Patterns from Kernel PCA and Bagging." IIE Transactions, Vol.46, Iss.5, 2014

[19]Ho, Tin Kam (1995). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995.vol.11 pp. 278–282.

[20] Dan Gao1,2, Yan-Xia Zhang1 and Yong-Heng Zhao1

 

"Random forest algorithm for classification of multiwavelength data". Research in Astron. Astrophys. 2009 Vol. 9 No. 2, 220–226
[21]Kleinberg, Eugene (1996). "An Overtraining-Resistant Stochastic Modeling Method for Pattern Recognition". Annals of Statistics. 24 (6):2319–2349. MR 1425956.
[22]Kleinberg, Eugene (2000). "On the Algorithmic Implementation of Stochastic Discrimination". IEEE Transactions on PAMI. 22 (5).
[23] Kleinberg, Eugene (2000). "On the Algorithmic Implementation of Stochastic Discrimination" (PDF). IEEE Transactions on PAMI. 22 (5).
[24] Breiman, Leo (2001). "Random Forests". Machine Learning. 45 (1): 5–32.
[25]Liaw, Andy (16 October 2012). "Documentation for R package Random Forest" (PDF). Retrieved 15 March 2013.
[26] ] Y. Lihua, D. Qi, and G. Yanjun, "Study on KNN Text Categorization Algorithm", Micro Computer Information, Vol.21, pp. 269-271, 2006

**Authors Profile**

*Ms. Heavenpreet Kaur* did her Diploma in CSE in 2011 and did B.Tech in CSE 2015. Currently she is pursuing her M.Tech in big data from IKG Punjab Technical University, Kapurthala, India.

*Mr. Dinesh Gupta*, did his M. Tech in IT from Department of CSE GNDU Amritsar, India. Currently he is pursuing his Ph. D in CSE. He has more than 8 years of experience in teaching. Currently he is working as Assistant Professor in department of CSE, IKGPTU, India. He has more than 8 publications in leading research Journal.

 