Engineering Universe for scientific Research and management
ISSN (Online): 2319-3069
Vol. 11 Issue V
May 2019

Prediction of Heart Disease Using 2-Tier SVM Data Mining Algorithm

Leena Sarvaiya[1], Prof.Himanshu Yadav[2], Prof. Chetan Agrawal[3]
[1,2,3] Department of Computer Science Engineering, RITS, Bhopal (MP), India

ABSTRACT

Data Mining is a field which extensively used in medical industry for extracting the useful information of various diseases. Heart disease is the major problem which occurs most frequently in human being. During the prediction of heart disease huge amounts of data generated which are too complex and voluminous to be processed and analyzed by traditional methods. By using data mining methodology it takes less time for the prediction of the disease with more accuracy. In this paper, proposes a hybrid SVM and KNN which can effectively diagnose and extract useful information from the outsized dataset. The simulation tool MATLAB is used for experimental analysis between existing methods and propose method using measuring parameter sensitivity, specificity and accuracy. The simulation of propose method gives better results than and can effectively diagnose the problem of heart disease.

Index Terms **- Heart Disease, Data Mining, Medical Industry, SVM, KNN, MATLAB**

1. INTRODUCTION

In the current society, most of the general population impacting heart disease. In this rapidly developing world people need to go ahead with a sumptuous life. With a particular true objective to get a lot of money and continue with a pleasant life they work like a machine. They never focus the organs of their body. Step by step the hazard factors are extended for heart disease. Depends on their working condition they change their nourishment propensities. It prompts blood pressure, sugar in energetic age. As a general rule people never direct a therapeutic authority. They are taking their own specific pharmaceutical. This kind of method drives indications. At introduce, a substantial number of people encounter the evil impacts of heart disease consistently. Heart disease is an essential explanation behind depressingness and mortality. As demonstrated by the World Health Organization, 12 million passing are realized by heart disease on the world consistently, 50 percent of which can be balanced by controlling risk factors.[2] Heart disease are foreseen that would be the crucial reason behind 35 to 60 percent of total demise expected worldwide by 2025.[1] Data mining is the technique toward isolating covered learning from data. It can reveal the cases and associations among tremendous measure of data in a single or a couple of datasets. Data mining techniques gives a client oriented approach to manage novel and covered cases in the information. In a manner of speaking Data mining is one of the methods for data divulgence for isolating comprehended cases from unfathomable, divided and disorderly information. It is a field with the crossroads of various controls that has brought true examination, bolster vector machine [11], KNN classifier [12] and database administration system together to address the issues. There are different hazard components which may improve the no. of patients of coronary heart disease.

.

1.1 The risk factor for heart disease [2]

- *Family history of heart disease*

Most people know that the heat disease can run in families. That if anybody has a family history of heart disease, he/she may be at greater risk for heart attack, stroke and other heard diseases.

- *Smoking*

Smoking is major cause of heart attack, stroke and other peripheral arterial disease. Nearly 40% of all people who die from smoking tobacco do so due of heart and blood vessel diseases. A smoker's risk of heart attack reduces rapidly after only one year of not smoking.

- *Cholesterol*

Abnormal levels of lipids (fats) in the blood are risk factor of heart diseases. Cholesterol is a soft, waxy substance found among the lipids in the bloodstream and in all the body's cells. High level of triglyceride (most common type of fat in body) combined with high levels of LDL (low density lipoprotein) cholesterol speed up atherosclerosis increasing the risk of heart diseases.

- *High blood pressure*

High blood pressure also known as HBP or hypertension is a widely misunderstood medical condition. High blood pressure increase the risk of the walls of our blood vessels walls becoming overstretched and injured. Also increase the risk of

having heart attack or stroke and of developing heart failure, kidney failure and peripheral vascular disease.

- *Obesity*

The term obesity is used to describe the health condition of anyone significantly above his or her ideal healthy weight. Being obese puts anybody at a higher risk for health problem such as heart disease, stroke, high blood pressure, diabetes and more.

- *Lack of physical exercise*

Lack of exercise is a risk factor for developing coronary artery disease (CAD). Lack of physical exercise increases the risk of CAD, because it also increases the risk for diabetes and high blood pressure.

1.2 Source of Heart Disease Dataset

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date.[3] The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2, 3, 4) from absence (value 0).
Only 14 attributes used:

1. #3 (age) 2. #4 (sex) 3. #9 (cp)
4. #10 (trestbps) 5. #12 (chol) 6. #16 (fbs)
7. #19 (restecg) 8. #32 (thalach) 9. #38 (exang)
10. #40 (oldpeak) 11. #41 (slope) 12. #44 (ca)
13. #51 (thal) 14. #58 (num) (the predicted attribute)

1.3 Types of Heart Disease

Heart disease is a broad term that includes all types of diseases affecting different components of the heart. Heart means 'cardio.' Therefore, all heart diseases belong to the category of cardiovascular diseases. Some types of Heart diseases are:[1]

- *Coronary heart disease*

It also known as coronary artery disease (CAD), it is the most common type of heart disease across the world. It is a condition in which plaque deposits block the coronary blood vessels leading to a reduced supply of blood and oxygen to the heart.

- *Congestive heart failure*

It is a condition where the heart cannot pump enough blood to the rest of the body. It is commonly known as heart failure.

- *Cardiomyopathy*

It is the weakening of the heart muscle or a change in the structure of the muscle due to inadequate heart pumping. Some of the common causes of cardiomyopathy are hypertension, alcohol consumption, viral infections, and genetic defects.

- *Congenital heart disease*

It also known as congenital heart defect, it refers to the formation of an abnormal heart due to a defect in the structure of the heart or its functioning. It is also a type of congenital disease.

- *Arrhythmias*

It is associated with a disorder in the rhythmic movement of the heartbeat. The heartbeat can be slow, fast, or irregular. These abnormal heartbeats are caused by a short circuit in the heart's electrical system.

- *Myocarditis*

It is an inflammation of the heart muscle usually caused by viral, fungal, and bacterial infections affecting the heart. It is an uncommon disease with few symptoms like joins pain, leg swelling or fever that cannot be directly related to the heart.

II. DATA MINING TECHNIQUES

Data Mining is main concerned with the analysis of data and Data Mining tools and techniques are used for finding patterns from the data set. The main objective of Data Mining is to find patterns automatically with minimal user input and efforts. Data Mining is a powerful tool capable of handling decision making and for forecasting future trends of market. Data Mining tools and techniques can be successfully applied in various fields in various forms. Many Organizations now start using Data Mining as a tool, to deal with the competitive environment for data analysis. By using Mining tools and techniques, various fields of business get benefit by easily **evaluate** various trends and pattern of market and to produce quick and effective market trend analysis. Data mining is very useful tool for the diagnosis of diseases. There are various techniques in data mining which helps in the prediction and extraction of useful information of heart patients such as clustering, classification, decision tree etc.

2.1 *Clustering*

Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. It helps users to understand the natural grouping or structure in a data set. Clustering is an unsupervised classification and has no predefined classes. They are used either as a stand-alone tool to get insight into data distribution or as a pre-processing step for other algorithms. Moreover, they are used for data compression, outlier detection, understand human concept formation. Some of the applications are Image processing, spatial data analysis and pattern recognition. Classification via Clustering is not performing well when compared to other two algorithms.[6]

2.2 Association

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in heart disease prediction as it tell us the relationship of different attributes used for analysis and sort out the patient with all the risk factor which are required for prediction of disease.[2]

2.3 Decision Trees

Decision Trees (DTs) are a non-parametric supervised learning method used for classification. The main aim is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The structure of decision tree is in the form of a tree. Decision trees classify instances by starting at the root of the tree and moving through it until a leaf node. Decision trees are commonly used in operations research, mainly in decision analysis. Some of the advantages are they can be easily understand and interpret, robust, perform well with large datasets, able to handle both numerical and categorical data. Decision-tree learners can create over-complex trees that do not generalize well from the training data is one the limitation. [6]

2.4 Prediction

The prediction as it name implied is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. For instance, prediction analysis technique can be used in sale to predict profit for the future if we consider sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction. [2]

III. RELATED WORK

Dewan and Sharma [4] developed a prototype which can find out and extract unknown knowledge (patterns and relations) related with heart disease from a past heart disease database record. It can resolve complicated queries for detecting heart disease and hence assist medical practitioners to make smart clinical decisions which traditional decision support systems were not able to. By providing proficient treatments, it can help to decrease costs of treatment.

Ranganatha et al. [5] stored medical information of patients who come for hospitalization for heart disease and algorithms are run on that information and result will be provided in the form of user understandable words and graph. When very large data sets are present, data mining algorithms (here considering only ID3 and Naïve Bayesian algorithms) are used. ID3 outputs the result in the form of decision tree which can be easily understood. Naïve Bayesian predicts the chances of heart disease based on conditions given.

Venkatalakshmi and Shivashankar [6] design and develop diagnosis and prediction system for heart diseases based on predictive mining. Number of experiments has been conducted to compare the performance of various predictive data mining techniques including Decision tree and Naïve Bayes algorithms. In this proposed work, a 13 attribute structured clinical database from UCI Machine Learning Repository has been used as a source data. Decision tree and Naive Bayes have been applied and their performance on diagnosis has been compared. Naïve Bayes outperforms when compared to Decision tree.

Sen et al. [7] proposed a layered neuro-fuzzy approach to predict occurrences of coronary heart disease simulated in MATLAB tool. The implementation of the neuro-fuzzy integrated approach produced an error rate very low and a high work efficiency in performing analysis for coronary heart disease occurrences.

Chandra et al.[8] proposed a new approach for association rule mining based on sequence number and clustering transactional data set for heart disease predictions. The implementation of the proposed approach was implemented in C programming language and reduced main memory requirement by considering a small cluster at a time in order to be considered scalable and efficient.

Deekshatulu et al. [9] created class association rules using feature subset selection to predict a model for heart disease. Association rule determines relations amongst attributes values and classification predicts the class in the patient dataset. Feature selection measures such as genetic search determines attributes which contribute towards the prediction of heart diseases.

Agrawal et al. [10] implemented a hybrid system that uses global optimization benefit of genetic algorithm for initialization of neural network weights. The prediction of the heart disease is based on risk factors such as age, family history, diabetes, hypertension, high cholesterol, smoking, alcohol intake and obesity.

IV. PROPOSED METHODOLOGY

Here heart dieses dataset [3] and pre-process dataset by converting into xlxs format. The Entropy based feature selection is to be done by finding Entropy and information gain of each attribute separately. Information gain measure is to be calculated using decision tree algorithm. Entropy is commonly used to calculate impurity. Information content is more when impurity is higher. Information Gain is metric for how well one attribute Ai classifies the training data. Decision Tree recursively partitions the training set Parameter which best classifies data is Entropy (H). Entropy is a good measure of the information carried by an ensemble of events. Entropy of set S is denoted by H(S). If S=Sample of n training events and Pi is the probability of occurrence of event, then entropy is given by:

$$H(S) = \sum_{i=1}^{n} - P_i \log_2 P_i$$

For each attribute calculate the information gain. Information gain is a statistical quantity measuring how well an attribute classifies the data. We have calculated the information gain (Gain(S, Ai)) for each attribute using Algorithmic Approach and in the end attribute with the highest information gain will be chosen for decision-making. *Sv* is the subset of *S* for which attribute *A* has value *v*.

$$Gain(S, A_i) = H(S) - \sum_{v \epsilon Values(Ai)} P(A_i=v)\, H(S_v)$$

Information gain is our metric for how well one attribute Ai classifies the training data.
The process is normalized using min-max algorithm to provide linear transformation on original range of data. This specifically fit the data by finding new range from an existing one range. Here we use new approach of level 2 classification using SVM in level 1 classification and KNN in level 2 classifications. Start level 1 classification process on training and test data using Support vector Machine classification. After prediction, now predicted classes will passed for training. On training data KNN based Level 2 classification will be applied. Classified result of new classes goal0, goal1, goal2, goal3 etc. will be calculated. Performance and prediction of dieses will be predicted by calculating standard parameters like accuracy, sensitivity and specificity of all four classes. All these parameters are examined and compared with previous SVM and KNN based algorithm.

4.1 Support Vector Machine

Support Vector Machine (SVM) is a category of universal feed forward networks like Radial-basis function networks, pioneered by Vapnik. SVM can be used for pattern classification and nonlinear regression. More precisely, the support vector machine is an approximate implementation of the method of structural risk minimization. This principle is based on the fact the error rate of a learning machine on test data is bounded by the sum of the training-error rate and term that depends on the Vapnik-Chervonenkis (VC) dimension. The support vector machine can provide good generalization performance on pattern classification problem [11]. Optimal Hyperplane for patterns: Consider the training sample $\{(x_i, y_i)\}_{i=1}^{N}$ where x_i is the input pattern for the i^{th} instance and yi is the corresponding target output. With pattern represented by the subset yi= +1 and the pattern represented by the subset yi= -1 are linearly separable. The equation in the form of a hyperplane that does the separation is

$$w^T x + b = 0 \quad (1)$$

Where x is an input vector, w is an adjustable weight vector, and b is a bias. Thus,

$$w^T x_i + b \geq 0 \text{ for } yi = +1 \quad (2)$$

$$w^T xi + b < 0 \text{ for } yi = -1 \quad (3)$$

For a given weight vector w and a bias b, the separation between the hyperplane defined in eq. 1 and closest data point is called the margin of separation, denoted by ρ as shown in figure 1, the geometric construction of an optimal hyperplane for a two- dimensional input space.

The discriminant function gives an algebraic measure of the distance from x to the optimal hyperplane for the optimum values of the weight vector and bias, respectively.

$$g(x) = w_o T_x + b_o \quad (4)$$

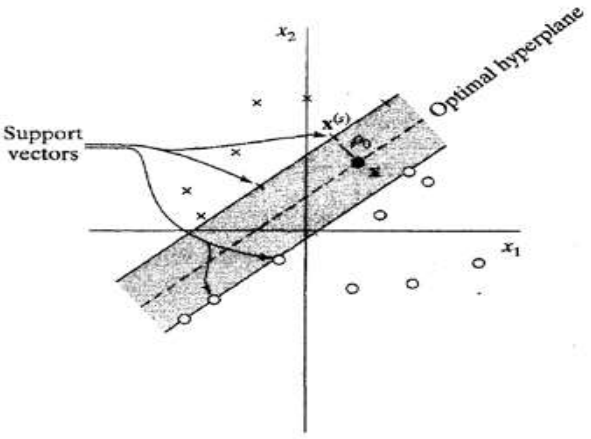

Fig 1: Optimal Hyperplane for a two tier input space

4.2 K nearest neighbor classifier

K nearest neighbor (KNN) is a simple algorithm, which stores all cases and classifies new cases based on similarity measure. KNN algorithm also called as 1) case based reasoning 2) k nearest neighbor 3)example based reasoning 4) instance based learning 5) memory based reasoning 6) lazy learning [4].KNN algorithms have been used since 1970 in many applications like statistical estimation and pattern recognition etc.KNN is a non parametric classification method which is broadly classified into two types1) structure less NN techniques 2) structure based NN techniques. In structure less NN techniques whole data is classified into training and test sample data. From training point to sample point distance is evaluated, and the point with lowest distance is called nearest neighbor. Structure based NN techniques are based on structures of data like orthogonal structure tree (OST), ball tree, k-d tree, axis tree, nearest future line and central line [12].Nearest neighbor classification is used mainly when all the attributes are continuous. Simple K nearest neighbor algorithm is shown in figure 2.

Steps 1) find the K training instances which are closest to unknown instance
Step2) pick the most commonly occurring classification for these K instances

Fig 2: K nearest neighbor algorithm

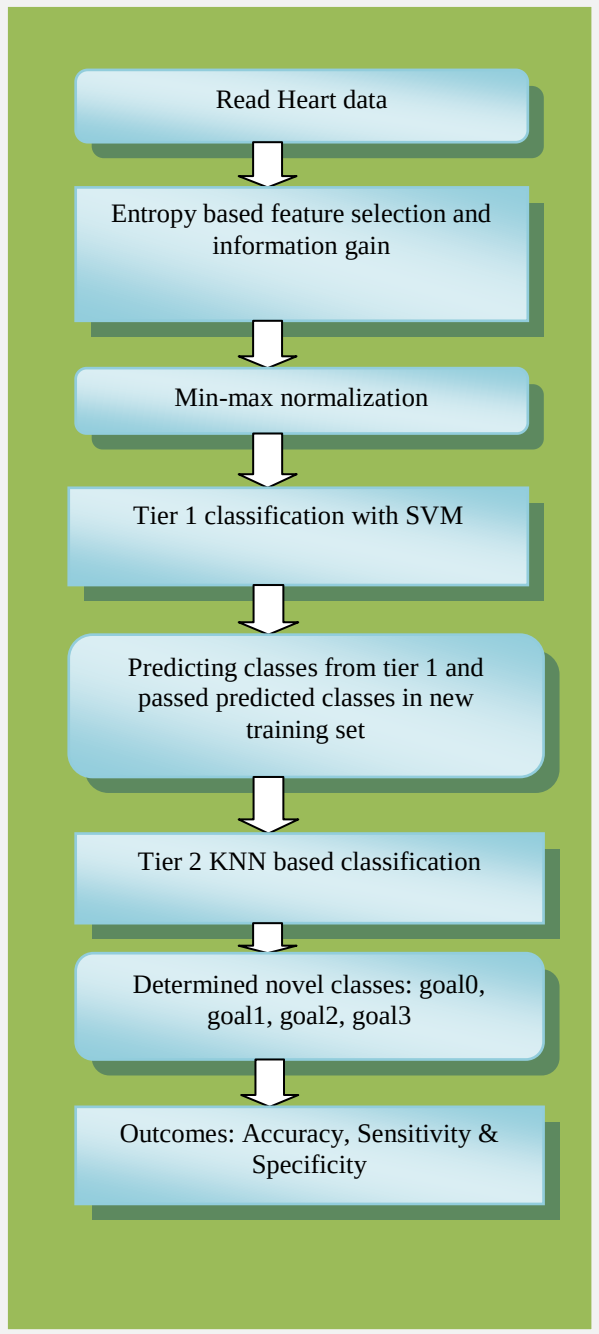

Fig 3: Block Diagram of proposed

V. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we present the results from our extensive experiments to compare the performance of M.KNN, SVM and 2tier proposed method on real health care data from a Chinese city. All the experiments are conducted on the MATLAB platform, which includes three Intel 3.4 GHz machines, each running on 16GB RAM.

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1, 2, 3, 4) from absence (value 0). Only 14 attributes used [3].

5.2 GUI Environment

This section shows the main GUI environment of the proposed methodology and probability entropy gain process of it.

Fig. 4: Main GUI of the Proposed Methodology

5.3 Result Analysis

The result analysis of the proposed work is performed using the accuracy and specificity parameter and for the processed attributes confusion matrix is formed which is shown below:

Table 1

Confusion Matrix for SVM	
Confusion Matrix of 'c2'	
98	2
35	16
0	0
Confusion Matrix of 'c1'	
60	15
64	12
0	0
Confusion Matrix of 'c3'	
95	4
39	13
0	0
Confusion Matrix of 'c4'	
136	4
9	2
1	0

Table 2

Confusion Matrix for Proposed Methodology	
Confusion Matrix of 'c2'	
246	21
14	22
0	0
Confusion Matrix of 'c1'	
233	15
22	33
0	0
Confusion Matrix of 'c3'	
251	17
14	21
0	0
Confusion Matrix of 'c4'	
281	9
6	7
0	0

Table 3

Confusion Matrix for Proposed 2tier	
Confusion Matrix of 'c2'	
246	9
14	20
0	0
Confusion Matrix of 'c1'	
233	27
22	40
0	0
Confusion Matrix of 'c3'	
251	9
14	19
0	0
Confusion Matrix of 'c4'	
281	3
6	7
0	0

5.3.1 Accuracy Analysis

The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

Accuracy=(TP+TN)/TP+TN+FP+FN

For this parameter the comparison between SVM, M.KNN and 2tier proposed method is perform in which it is found that the accuracy rate of SVM is about 75%, M.KNN is 89% and our method is about 90% which means our method generates better accuracy rate than the existing SVM method.

Table 4: Accuracy result analysis of the proposed 2tier method

Accuracy			
	SVM	M.KNN	2-tier
goal2	0.754967	0.88488	0.9174
goal1	0.476821	0.877888	0.861386
goal3	0.715232	0.89769	0.917492
goal4	0.913907	0.950495	0.970297

Fig. 5: Accuracy graph between SVM, M.KMM and 2tier proposed Method

5.3.2 Specificity Analysis

The sensitivity of a test is its ability to determine the patient cases correctly. To estimate it, we should calculate the proportion of true positive in patient cases. Mathematically, this can be stated as:

Sensitivity=TP/TP+FN

For this parameter the comparison between SVM and proposed method is perform in which it is found that the accuracy rate of SVM is about 60%, M.KNN is about 54 % and our method 2tier is about 59% which means our method generates better specificity rate than the existing SVM method and M.KNN.

Table 5: Specificity result analysis of the SVM, M.KNN and 2 tier Proposed method

Specificity			
	SVM	M.KNN	2-tier
goal2	0.888889	0.511628	0.6896
goal1	0.44444	0.6875	0.59701
goal3	0.764706	0.552632	0.6785
goal4	0.33333	0.4375	0.7

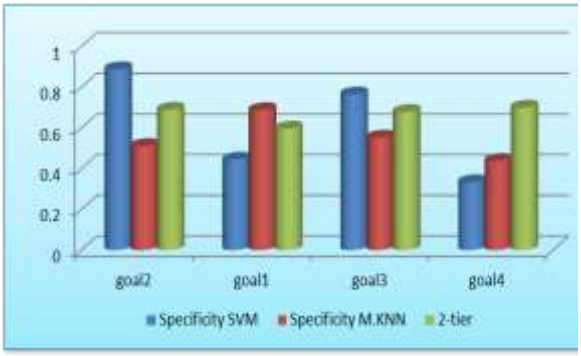

Fig. 6: Specificity graph between SVM, M.KNN and 2 tier Proposed method

5.3.3 Senstivity Analysis

Sensitivity (also called the true positive rate, the recall, or probability of detection in some fields) measures the proportion of positives that are correctly identified as such (i.e. the percentage of sick people who are correctly identified as having the condition).

Specificity=TN/TN+FP

For this parameter the comparison between SVM and proposed method is perform in which it is found that the accuracy rate of SVM is about 71%, M.KNN is about 94% and our method 2 tier is about 95% which means our method generates better specificity rate than the existing SVM method.

Table 6: Sensitivity result analysis of the SVM, M.KNN and 2 tier Proposed method

Sensitivity			
	SVM	M.KNN	2-tier
goal2	0.736842	0.946154	0.9416
goal1	0.483871	0.913725	0.93644
goal3	0.708955	0.94717	0.94181
goal4	0.937931	0.979094	0.979522

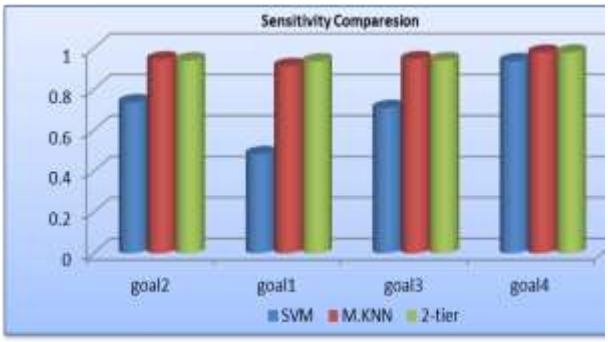

Fig. 7: Senstivity graph between SVM, M.KNN and 2 tier Proposed method

VI. CONCLUSION

This article presented a novel approach 2tier for classifying heart disease. As a way to validate the proposed method, this is tested on machine learning data sets taken from UCI repository. We have selected only 14 dataset among 76 dataset for the prediction of heart disease patient. This proposed prediction model helps the doctors in proficient heart disease diagnosis process with fewer attributes. This disease is mostly found in India and in Andhra Pradesh. The analysis of proposed method is performing using sensitivity, specificity and accuracy. The simulation result of accuracy parameter of our proposed method is about 95% which too much greater than SVM method and some more than M.KNN.

REFERENCES

[1] K. Manimekalai "Prediction of Heart Diseases using Data Mining Techniques", International Journal of Innovative Research in Computer and Communication Engineering Vol. 4, Issue 2, February 2016, ISSN(Online): 2320-9801

[2] Beant Kaur,Williamjeet Singh "Review on Heart Disease Prediction System using Data Mining Techniques", International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 2 Issue: 10, ISSN: 2321-8169 pp: 3003 – 3008.

[3] Blake, C.L., Mertz, C.J.: "UCI MachineLearningDataset", http://mlearn.ics.uci.edu/databases/heartdisease, 2004.

[4] Ankita Dewan, Meghna Sharma "Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification", In proceeding of IEEE 2015.

[5] Ranganatha S., Pooja Raj H.R., Anusha C., Vinay S.K. "Medical Data Mining and Analysis for Heart Disease Dataset using Classification Techniques", In proceeding of IEEE 2013.

[6] B.Venkatalakshmi, M.V Shivsankar "Heart Disease Diagnosis Using Predictive Data mining", International Conference on Innovations in Engineering and Technology (ICIET'14) On 21st&22ndMarch, Volume 3, Special Issue 3. In proceeding of IJIRSET.

[7] A. K. Sen, S. B. Patel, and D. P. 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*, vol. 2, no. 9, pp. 1663–1671, 2013.

[8] M. Jabbar, P. Chandra, and B. Deekshatulu, "Cluster Based Association Rule Mining For heart attack prediction," Journal of Theoretical &

Applied Information Technology, vol. 32, no. 2, pp. 196–201, 2011.

[9] P. Chandra, M. . Jabbar, and B. . Deekshatulu, "Prediction of Risk Score for Heart Disease using Associative Classification and Hybrid Feature Subset Selection," in *12th International Conference on Intelligent Systems Design and* Applications (ISDA), 2012, pp. 628–634.

[10] S. U. Amin, K. Agarwal, and R. Beg, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors," in Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013), 2013, no. Ict, pp. 1227–1231.

[11] Christopher J.C. Burges. "A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery" Springer, 2(2), pp.121-167, 1998.

[12] Nitin Bhatia ,vandana" Survey on nearest neighbor techniques"IJCSIS,Vol 80,no 2(2010)