International Journal of Engineering & Technology, 7 (4.5) (2018) 292-296

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Classification of Heart Disease Hungarian Data Using Entropy, Knnga Based Classifier and Optimizer

Shweta Gupta*

**Department of Computer Science & Application*
AKS University Satna, India
**Corresponding author E-mail shwetaguptacs03@gmail.com*

Abstract

To mine the useful information from massive medical databases data mining plays as imperative role. In data mining classification (supervised learning) which can be used to design model by describing significant data classed, where class attribute is involved in the construction of the classifier. In this work, we propose a methodology in which uses KNN classifier. It is simple, popular, more efficient and proficient algorithm for pattern recognition. The samples of the medical databases are classified on the basis of nearest neighbor in which medical database are massively found in nature and contains irrelevant and redundant attributes. The only KNN classifier produce less accurate results that is why we use hybrid approach of KNN and genetic algorithm (GA) to obtain more accurate results. To evaluate the performance of the proposed approach Hungarian dataset (UCI learning) is used to classify the attributes of heart disease. The genetic algorithm performs global research on complex large and multimodal landscapes which provide minimal solutions or search space. The experimental outcomes of accuracy parameter of proposed approach give more accurate and efficient results than the existing approach.

Keywords: *Heart disease, neural network, support vector machine, genetic algorithm, k nearest neighbors.*

1. Introduction

Heart disease is nothing but the class of diseases that involve the heart or blood vessels (arteries and veins). Today most nations confront high and developing rates of coronary illness and it has turned into a main source of weakening and demise worldwide in men and ladies over age sixty-five and nowadays in several countries heart disease is seen as a "second epidemic," replacing infectious diseases as the main source of death [1]. Most countries face high and increasing rates of heart disease or Cardiovascular Disease. Even though, modern medicine is generating huge amount of data every day, little has been done to use this available data to solve the challenges that face a successful interpretation of heart disease examination results. Data mining is a knowledge discovery technique to analyze data and encapsulate it into useful information. The current analysis intends to predict the chance of obtaining heart disease given patient knowledge set. Predictions and descriptions are principal goals of data mining, in practice. Prediction in data mining involves attributes or variables in the data set to find unknown or future state values of other attributes. the information to be taken by humans. the aim of predictions in data processing is to assist discover trends in patient knowledge so as to boost their health[1]. Because of amendment in life designs in developing countries, like African country, Cardio Vascular Disease (CVD) has become a number one reason for deaths CVD is projected to be one largest killer worldwide accounting for all deaths. a shot to use information, expertise and clinical screening of patients to diagnose or acknowledge heart attacks is thought to be a wanted chance. Within the health sectors data processing plays a vital role to predict diseases. The prophetic finish of the analysis may be a data processing model and figure 1 show the stages concerned in data processing.

1.1. Applications of Data Mining To Healthcare Data

Data mining scholars have long studied the appliance of tools and instrumentality in up the method of information analysis in giant and complicated datasets. Adopting data processing techniques within the medication field is of high importance in diagnosis, predicting and deeply understanding of care knowledge. These applications embody treatment centers analysis geared toward up treatment policies and hindrance of any mistake in hospitals, early identification of diseases, hindrance of diseases and hospital death reduction.

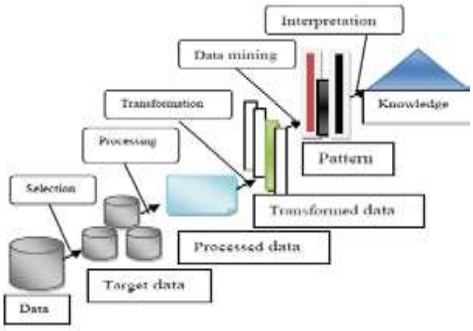

Fig. 1: Steps involved in data mining

Heart, specialist's record and store large amounts of patients' data. This provides a great opportunity for extracting a valuable knowledge from such datasets. Researcher's area unit adopting applied mathematics approaches still as data processing techniques to assist treatment and health care specialists diagnose and verify heart condition risk factors in patients. Applied mathematics analyses have known variety of risk factors for heart diseases

together with age, pressure level, smoking, total steroid alcohol, diabetes, and high blood pressure, heart condition background in family, fat and lack of physical activity. the attention of heart condition risk factors assists treatment and health care specialists to spot patients United Nations agency area unit subject to high risk factors.

2. Factor Affecting the Heart

The circumstances or habits that make a person more likely to develop a disease are Risk factors. They can also boost the probability of an existing disease will get worse[2].

2.1. Controllable Risk Factors

Smoking: The chemicals in tobacco smoke promote the development of blood clots and increase the cause heart attacks by building-up of plaque in artery walls.

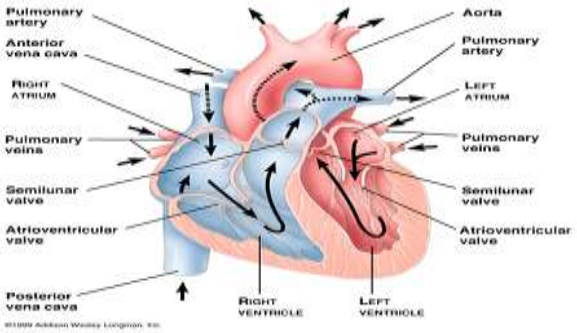

Fig. 2: Structure of heart

Weight: If body pound increases, the risk of heart disease also rises. This is especially factual for people who carry extra body fat around the waist. To reduce the risk of heart disease numerous dietary factors that can be used.
Cholesterol: Excessive cholesterol in the blood building up in the walls of the arteries can cause a process called atherosclerosis, a form of heart disease.
Diabetes: Diabetes can cause heart disease by growing the risk of high blood pressure and high cholesterol in the blood. It promotes injury to the artery walls and formation of blood clots.
Blood pressure: Blood pressure is the force of the blood against the inner walls of the blood vessels, generated when the heart pumps blood. When a person has hypertension, the arteries are under increased pressure and the heart has to pump harder, which may lead to injury of the artery walls, atherosclerosis, and coronary heart disease.

2.2. Uncontrollable Risk Factors

Age: Heart Related disease usually occurs in women after menopause and in men above the age of 40.
Sex: Men have got more risk of heart attack than women, and men generally suffer from heart attacks at earlier ages.
Family history: For the person who is having a close relative who had heart attack may be at risk of heart disease.

Fig. 3: Causes of Heart Attack

3. Related Work

Author/ Researcers	Description
G.Vaishali, V.Kalaivani[3]	Developed centralized patient observation system victimization massive information. Within the planned system, giant set of medical records area unit taken as input. From this medical dataset, it's aimed to extract the required data from the record of heart patients victimization map scale back technique. Heart disease may be a major pathological state and it's the leading causes of death throughout the globe. Early detection of Heart disease has become a very important issue within the medical analysis fields. For Heart disease detection, some options area unit analyzed like RR interval, QRS interval and QT interval. The classification method states whether or not the patient is traditional or abnormal and within the noticeion step victimization map scale back technique to detect the disease and scale back the dataset.
Ankita Dewan, Meghna Sharma[4]	Developed an image which may verify and extract unknown data (patterns and relations) connected with Heart disease from a past Heart disease info record. It will resolve difficult queries for detection Heart disease and thence assist medical practitioners to form sensible clinical choices that ancient decision support systems weren't able to. By providing adept treatments, it will facilitate to decrease prices of treatment.
B. Venkatalakshmi, M.V Shivsankar[5]	Design and develop identification and prediction system for heart diseases supported prognosticative mining. variety of experiments has been conducted to match the performance of varied prognosticative data processing techniques as well as call tree and Naïve Thomas Bayes algorithms. during this planned work, a thirteen attribute structured clinical info from UCI Machine Learning Repository has been used as a supply information. decision tree and Naive Thomas Bayes are applied and their performance on identification has been compared. Naïve Thomas Bayes outperforms when put next to Decision tree.
S. U. Amin, K. Agarwal, and R. Beg,[6]	Implemented a hybrid system that uses world optimization advantage of genetic algorithmic program for formatting of neural network weights. The prediction of the center disease relies on risk factors like age, case history, diabetes, high blood pressure, high cholesterol, smoking, alcohol intake and fat.
A. K. Sen, S. B. Patel, and D. P. Shukla[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 made a slip rate terribly low and a high work potency in activity analysis for coronary Heart disease occurrences.
B.Venkatalakshmi, M.V Shivsankar[8]	This project intends to design and develop diagnosis and prediction system for heart diseases supported prognosticative mining. Variety of experiments has been conducted to match the performance of varied prognosticative data processing techniques as well as decision tree and Naïve Thomas Bayes algorithms. during this planned work, a thirteen attribute structured clinical info from UCI Machine Learning Repository has been used as a supply information. decision tree and Naive Thomas Bayes are applied and their performance on identification has been compared. Naïve Thomas Bayes outperforms when put next to Decision tree

4. Data Mining Techniques

Researchers have utilized completely different data processing techniques to help out specialist and physician diagnose heart condition[9]. Some techniques square measure additional common like Naïve Bayes, decision tree and K-nearest neighbor. However, there square measure different classification-based data processing techniques like kernel density, neural network, bagging algorithm, sequential minimal optimization, direct Kernel self-organizing map and support vector machine. Successive section in short explains those techniques that were employed in this study.

4.1. Decision Tree

There are different classes of decision trees. They only differ in the numerical model they use to select the class of attribute during rule extraction. Gain ratio decision tree is the most common, successful type [10]. It is an association between entropy (information gain) and classified information. In entropy procedure, the attribute which minimizes entropy and maximizes information gain is chosen as the tree root. To choose tree root, it is first essential to estimate the information gain of each attribute. Then, the attribute maximizing information gain should be chosen. Information gain, or entropy[11].

$$E = -\sum_{i=1}^{k} p_i log_2^{p_i} \quad (1)$$

Where k is the number of response variable classes, pi is the ratio of the number of the ith class events to total number of samples (occurrence probability of i)

4.2. Bayesian Network

Bayesian network is a statistical technique predicting the membership class of the studied sample using the probability theory. Bayesian network practices classification process in accordance with Bayes' theorem. It assumes that the influence of the value of a theorem on a class is independent from the influence of other attributes. This assumption is called "class conditional independence". This assumption was made to shorten engaged calculation and this is why it was named "Naïve", i.e., simple. This technique calculates the prior probability of the response variable and the conditional probability of other variables. The prior and conditional probabilities of the initial training are calculated. Then, for every test dataset sample, the probability of the occurrence (presence) of each case of response variable is calculated. Afterwards, the response variable with the highest occurrence probability is selected. The probability of test sample for the response variable value is[12].

$$P(v = c_i = P(c_i) = \sum_{j=1}^{n} P(a_j = v_j | class = c_i) \quad (2)$$

Where V, ci, aj and vj are test sample, response variable value, data attribute and the test sample value, respectively.

4.3. Support Vector Machine

Given availability of support vectors, Support Vector Machine (SVM) is the boundary determining the best data classification and separation. In SVM, only those data lying inside support vectors are used as the base data for machine and building a model. This means that this algorithm is not sensitive to other data. It aims to find the best data boundary with the farthest possible distance from all classes (their support vectors). SVM transfers data to a new space with respect to their predetermined classes so that data can be classified and separated linearly (using hyperplanes). Then, it searches for support lines (or support planes in multi-dimensional space) and tries to determine the equation of a straight line that maximizes the distance between each two classes. Each support vector is characterized with an equation describing the boundary line of each class.

5. Proposed Methodology

In this segment of the analysis work, we tend to use KNN-GA classifier for the recognition of heart patient. The outline of the proposed methodology and its formula is described below:

5.1. K-Nearest Neighbour

K nearest neighbor(KNN) may be a straightforward algorithmic program, that stores all cases and classify new cases supported similarity live. KNN algorithmic program conjointly referred to as 1) case based mostly reasoning 2) k nearest neighbor 3)example based reasoning 4) instance based learning 5) memory based reasoning 6) lazy learning [14]. KNN algorithms are used since 1970 in several applications like applied math estimation and pattern recognition etc.KNN may be a non constant quantity classification methodology that is generally classified into 2 types 1) structure less NN techniques 2) structure primarily based NN techniques. In structure less NN techniques whole knowledge is assessed into coaching and check sample knowledge. From coaching purpose to sample purpose distance is evaluated, and also the purpose with lowest distance is termed nearest neighbor. Structure primarily based NN techniques area unit supported structures of knowledge like orthogonal structure tree (OST), ball tree, k-d tree, axis tree, nearest future line and central line[15]. Nearest neighbor classification is employed in the main once all the attributes area unit continuous.

Steps 1) realize the K coaching instances that area unit nighest to unknown instance
Steps 2) decide the foremost usually occurring classification for these K instances

KNN is employed in several applications like 1) classification and interpretation 2) drawback resolution 3) operate learning and teaching and coaching. KNN suffers from the subsequent drawbacks 1) low potency 2) dependency on the choice of excellent values for k.

5.2. Genetic Algorithm

In the area of AI, a genetic algorithmic program (GA) could be a search heuristic that imitates the method of natural evolution. This heuristic is habitually wont to produce accommodating solutions to optimisation and search issues. Genetic algorithms belong to the larger category of evolutionary algorithms (EA) that produce optimized solutions victimization techniques galvanized by natural evolution, like inheritance, mutation, selection, and crossover. A classic genetic algorithmic program requires:

1. A genetic illustration of the solution domain,
2. A fitness performs to judge the solution domain.

A standard illustration of solution is as an array of bits. The fitness perform is outlined over the genetic illustration and measures the standard of the delineated solution. The fitness perform is usually downside dependent. At first several individual solutions area unit (usually) haphazardly generated to make associate initial population. Throughout every consecutive generation, a proportion of the prevailing population is chosen to breed a brand new generation. Individual solutions area unit chosen through a fitness-based method, wherever fitter solutions (as measured by a fitness function) area unit usually additional seemingly to be chosen. Future step is to come up with a second generation population of

solutions from those chosen through genetic operators: crossover (also referred to as recombination) and mutation.

Proposed Steps
Step 1: Make X1 reduced datasets from a database.
Step 2: Set a learning algorithm to individual pattern for test dataset.
Step 3: Set a learning algorithm to individual pattern training dataset.
svmStruct = knnclassify(X1(train(:,i1),:),groups(train(:,1))
Step 4: Object with unknown found to do with each of the X1 classifiers predictions.
Step 5: Select the most repeatedly predicted samples.

KNNGA steps:
Step1: Initialize population = *from normalized_data*
Step2: Apply genetic search into selected dataset
Step3: Apply KNN classifier for testing of all classes (goal) data which are classified or misclassified data.
Step4: Each attribute will organize as per their ranks.
Step5: Higher ranked attribute will select.
Step6: Apply KNNGA () on the each five subset of the attributes for enhance the accuracy level.
Step7:If knnga_classifier(class_knn)>knn_classifer(class_knn)
data_class = class_knn;
else
data_class = class_knnga;
Step8: Perform the reproduction (if needed while iteration then go to step 9 else step 11)
Step9: Apply crossover operator
Step10: Perform mutation then produce new population *X'1*
Step11: Calculate the local maxima for each category.
Repeat the steps till iterations ane not finished
Step12: For each test *X'1*, start all trained base models then prediction of result by combining of all trained models, and separate the misclassified by optimized knnGA.

Here block diagram exposed in figure 4 that the working of proposed approach, where at initial state health care dataset is preferred for the processing, then into next stage entire dataset is logically separate for the moment due to it is containing string fields in addition to numeric fields, so in the designing approach they developed separate approach for string and numeric data.
Pre-processing: It converts the data which is additional consistent for unsupervised learning by deleting the labels from the dataset.
Data fraction: Preprocessed data are used to partition into training & testing sets samples.
Detection of Normal: In this step normal data is separated from the training data sample, here training process is complete by training and normalize using min-max.
And if the normal class has been easily detected then its goes to the separately normal class otherwise if not detected then it will go to the KNN-GA classifier. And in this process each class has been accurately predicted with their own characteristics, after successful prediction the result analysis approach follows for the detected intrusions.

6. Experimental Result

In this segment, we present the outcome from our extensive experiment to evaluate the performance of MKNN, SVM and 2tier proposed method on real health care data from a Chinese city. Every the experiments are conducted on the MATLAB platform, which includes three Intel 3.4GHz machines, each executing 4GB RAM.
This database involves 76 attributes; with the exception of every published experiment refer to utilizing a subset of 14 of them. Specifically, the Hungarian database is the special case that has been used by ML analysts to this date. The "class" field refers to the presence of heart disease in the patient. It is number esteem from 0 (no presence) to 4. Experiments with the Hungarian database have focused on just conceive to completely different presence (values 1, 2, 3, 4) from absence (value 0). Only 14 attributes used [16].

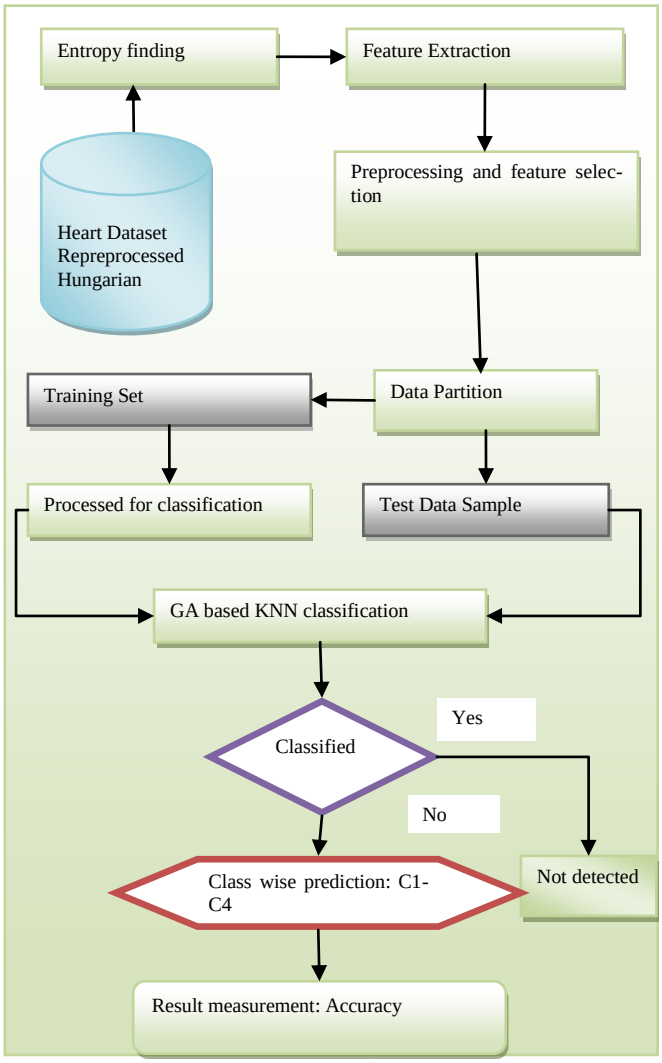

Fig. 4: Block diagram of proposed work

6.1. Heart Disease Dataset

This file involves 76 attributes; with the exception of every published experiment refer to utilizing a subset of 14 of them. Specifically, the Hungarian database is the special case that has been used by ML analysts to this date[16]. The "class" field refers to the presence of heart disease in the patient. It is number esteemed from 0 (no presence) to 4. Experiments with the Hungarian database have focused on just conceive to completely different 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)

6.2. Result Analysis

Accuracy is the calculation of the discriminating results between the patients and healthy subject classes. If the results of classification do not provide correct discrimination between alternative states of health, then the accuracy is not significant while correct discrim-

ination provides high accuracy. In which each goal field of SVM method is compared with the each goal field of KNN-GA method. In SVM method there are four goals (goal1, goal2, goal3, goal4) and goal1 accuracy is 0.7208, goal2 accuracy is 0.8252, goal3 accuracy is 0.7298 and goal4 accuracy is 0.913907. In KNN-GA method (Proposed method) there are four goals (goal1, goal2, goal3, goal4) and goal1 accuracy is 0.9215, goal2 accuracy is 0.9546, goal3 accuracy is 0.9265 and goal4 accuracy is 0.9625.

$$\text{Accuracy rate} = \frac{\text{Sum of accuracy of all goals}}{\text{No. of goals}} \text{X } 100 \quad (3)$$

For this parameter the comparison between SVM, projected KNN-GA classifier method is perform in which it is found that the accuracy rate of SVM is about 77%, and our KNN-GA method is about 95% is exposed in table 1 and graph are described in figure 5, which means our method generates enhanced accuracy rate than the presented SVM method.

Table 1:: Accuracy analysis

Table1: Accuracy		
	SVM	**Proposed**
class2	0.8252	0.9546
class1	0.7208	0.9215
class3	0.7298	0.9265
class4	0.8278	0.9625

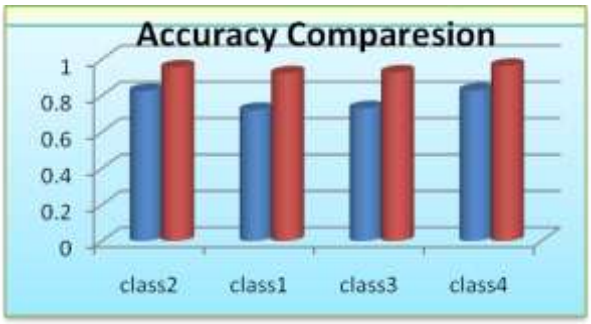

Fig. 5: Comparison graph of accuracy between SVM and Proposed method

7. Conclusion

The majority of the area of India is suffering from heart disease, so early prediction of this disease is much essential. Here we utilize KNN-GA classifier approach for the mining of helpful information from the massive amount of database. As an approach to approve our proposed strategy, the testing of heart disease machine learning dataset which is taken from UCI repository system. For the trial examination of the proposed approach utilizes just 14 dataset out of 76 dataset and the outcomes delivers by our approach gives more proficient than the current approach (SVM). This expectation show encourages the specialist to finding the coronary illness persistent by performing less perception.

References

[1] Aziz, N. Ismail, and F. Ahmad, "Mining Students' Academic Performance", Journal of Theoretical & Applied Information Technology, vol. 53, no. 3, 2013.

[2] S. Kiruthika Devi, S. Krishnapriya and Dristipona Kalita "Prediction of Heart Disease using Data Mining Techniques", Indian Journal of Science and Technology, Vol 9(39), DOI: 10.17485/ijst/2016/v9i39/102078, October 2016.

[3] G.Vaishali, V.Kalaivani "Big Data Analysis for Heart Disease Detection System Using Map Reduce Technique", In proceeding of IEEE, 2016.

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

[5] 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.

[6] 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.

[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] B.Venkatalakshmi, M.V Shivsankar "Heart Disease Diagnosis Using Predictive Data mining", International Journal of Innovative Research in Science, Engineering and Technology Volume 3, Special Issue 3, March 2014.

[9] Helma C, Gottmann E, Kramer S (2000) Knowledge discovery and data mining in toxicology. Statistical Methods in Medical Research 9: 329-358.

[10] Quinlan JR (1986) Decision trees and multi-valued attributes. In: Hayes, Michie D (eds.) Machine intelligence. Oxford University Press.

[11] Han J, Kamber M (2006) Data Mining Concepts and Techniques: Morgan Kaufmann Publishers.

[12] Bramer M (2007) Principles of data mining: Springer.

[13] K.Sudhakar , Dr. M. Manimekalai "Study of Heart Disease Prediction using Data Mining" International Journal of Advanced Research in Computer Science and Software Engineering 4(1), January - 2014, pp. 1157-1160.

[14] Shishir K. Shandilya, S. Jain, "Automatic opinion extraction from web documents", Proceeding of International Conference on Computer and Automation Engineering, pp. 351-355, 2009.

[15] Ashutosh Dubey and Shishir K. Shandilya,"Exploiting Need Of Data Mining Services in Mobile Computing Environments", Computational Intelligence and Communication Networks (CICN), 2010

[16] R. Chaure, Shishir K. Shandilya, "Firewall anomalies detection and removal techniques – A survey", International Journal of Emerging Technologies, Vol. 1(1), pp. 71–74, 2010

[17] A.K. Dubey, Shishir K. Shandilya, "A comprehensive survey of grid computing mechanism in J2ME for effective mobile computing techniques," Industrial and Information Systems (ICIIS), pp.207-212, 2010

[18] Shishir K. Shandilya, S. Jain, "Opinion Extraction & Classification of Reviews from Web Documents", Advance Computing Conference IEEE International, 2009.

[19] Asha Khilrani, Shishir K. Shandilya, "Implementation of User's Browse Log Monitoring Tool for Effective Web Usage Mining", International Journal of Computer Science and Information Technologies, Vol. 2 (3) pp. 1061-1064, 2011.

[20] Shishir K. Shandilya, S Jain, "Automatic Extraction and Classification of Opinions of Product Reviews from Web Documents", IUP Journal of Systems Management, 2011

[21] N Mishra, R Kumar, SK Shandilya, Credit Card Transaction Fraud Detection by using Hidden Markov Model, International Journal of Scientific Engineering and Technology, Volume No.1, Issue No.2 pp:139-142, 2277-1581, 2012

[22] Smita Shandilya, SK Shandilya, Tripta Thakur, Atulya K Nagar, Handbook of Research on Emerging Technologies for Electrical Power Planning, Analysis, and Optimization, 2016

[23] Shishir K. Shandilya, Smita Shandilya, Kusum Deep, Atulya K. Nagar, Handbook of Research on Soft Computing and Nature-Inspired Algorithms, 2017

[24] Shishir K. Shandilya, Suneet K. Gupta, A Comprehensive Survey on Author's Trait on Blog Data, International Journal of Advanced Engineering & Application, 2011

[25] S Shandilya, T Thakur, SK Shandilya, Transmission Network Expansion Planning Considering N-1 Contingency, Proceedings of International Conference on Control, Communication and Power Engineering, Elsevier, 2013

[26] Dr saed sayad,"University of toronto http://chem-eng.utoronto.ca/~data mining.

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