International Journal of All Research Education and Scientific Methods (IJARESM),
ISSN: 2455-6211, Volume 11, Issue 6, June-2023, Available online at: www.ijaresm.com

Cardiovascular Disease Prediction using Machine Learning Algorithm

**Ramya Kodali[1], Dipanjan Ghosh[2], Ushasree Tella[3], Hrithik Gajera[4],
Daipayan Hati[5], Kosuri Vaishnavi Chaitanya[6], Shreyansh Srikar Rao Polkampeta[7],
Lingala Niharika[8]**

---****************---

ABSTRACT

In the medical field, the diagnosis of heart disease is the most difficult task. The diagnosis of heart disease is difficult as a decision relied on grouping of large clinical and pathological data. Due to this complication, the interest increased in a significant amount between the researchers and clinical professionals about the efficient and accurate heart disease prediction. In case of heart disease, the correct diagnosis in early stage is important as time is the very important factor. Heart disease is the principal source of deaths widespread, and the prediction of Heart Disease is significant at an untimely phase. Machine learning in recent years has been the evolving, reliable and supporting tools in medical domain and has provided the greatest support for predicting disease with correct case of training and testing. The main idea behind this work is to study diverse prediction models for the heart disease and selecting important heart disease feature using Random Forests algorithm. Random Forests is the Supervised Machine Learning algorithm which has the high accuracy compared to other Supervised Machine Learning algorithms such as logistic regression etc. By using Random Forests algorithm we are going to predict if a person has heart disease or not.

INTRODUCTION

The heart is a kind of muscular organ which pumps blood into the body and is the central part of the body's cardiovascular system which also contains lungs. Cardiovascular system also comprises a network of blood vessels, for example, veins, arteries, and capillaries. These blood vessels deliver blood all over the body. Abnormalities in normal blood flow from the heart cause several types of heart diseases which are commonly known as cardiovascular diseases (CVD). Heart diseases are the main reasons for death worldwide. According to the survey of the World Health Organization (WHO), 17.5 million total global deaths occur because of heart attacks and strokes. More than 75% of deaths from cardiovascular diseases occur mostly in middle-income and low-income countries. Also, 80% of the deaths that occur due to CVDs are because of stroke and heart attack . Therefore, prediction of cardiac abnormalities at the early stage and tools for the prediction of heart diseases can save a lot of life and help doctors to design an effective treatment plan which ultimately reduces the mortality rate due to cardiovascular diseases.

Due to the development of advance healthcare systems, lots of patient data are nowadays available (i.e. Big Data in Electronic Health Record System) which can be used for designing predictive models for Cardiovascular diseases. Data mining or machine learning is a discovery method for analyzing big data from an assorted perspective and encapsulating it into useful information. "Data Mining is a non-trivial extraction of implicit, previously unknown and potentially useful information about data". Nowadays, a huge amount of data pertaining to disease diagnosis, patients etc. are generated by healthcare industries. Data mining provides a number of techniques which discover hidden patterns or similarities from data.

Therefore, in this paper, a machine learning algorithm is proposed for the implementation of a heart disease prediction system which was validated on two open access heart disease prediction datasets. Data mining is the computer based process of extracting useful information from enormous sets of databases. Data mining is most helpful in an explorative analysis because of nontrivial information from large volumes of evidence
.Medical data mining has great potential for exploring the cryptic patterns in the data sets of the clinical domain.

These patterns can be utilized for healthcare diagnosis. However, the available raw medical data are widely distributed, voluminous and heterogeneous in nature .This data needs to be collected in an organized form.

This collected data can be then integrated to form a medical information system. Data mining provides a user-oriented approach to novel and hidden patterns in the Data The data mining tools are useful for answering business questions and techniques for predicting the various diseases in the healthcare field. Disease prediction plays a significant role in data mining. This paper analyzes the heart disease predictions using classification algorithms. These invisible patterns can be utilized for health diagnosis in healthcare data.

Data mining technology affords an efficient approach to the latest and indefinite patterns in the data. The information which is identified can be used by the healthcare administrators to get better services. Heart disease was the most crucial reason for victims in the countries like India, United States. In this project we are predicting the heart disease using classification algorithms. Machine learning techniques like Classification algorithms such as Random forest, Logistic Regression are used to explore different kinds of heart based problems.

LITERATURE SURVEY

Machine Learning techniques are used to analyze and predict the medical data information resources. Diagnosis of heart disease is a significant and tedious task in medicine. The term Heart disease encompasses the various diseases that affect the heart. The exposure of heart disease from various factors or symptom is an issue which is not complimentary from false presumptions often accompanied by unpredictable effects. The data classification is based on Supervised Machine Learning algorithm which results in better accuracy. Here we are using the Random Forest as the training algorithm to train the heart disease dataset and to predict the heart disease. The results showed that the medicinal prescription and designed prediction system is capable of prophesying the heart attack successfully. Machine Learning techniques are used to indicate the early mortality by analyzing the heart disease patients and their clinical records (Richards, G. et al., 2001). (Sung, S.F. et al., 2015) have brought about the two Machine Learning techniques, k-nearest neighbor model and existing multi linear regression to predict the stroke severity index (SSI) of the patients. Their study show that k-nearest neighbor performed better than Multi Linear Regression model. (Arslan, A. K. et al., 2016) have suggested various Machine Learning techniques such as support vector machine (SVM), penalized logistic regression (PLR) to predict the heart stroke. Their results show that SVM produced the best performance in prediction when compared to other models. Boshra Brahmi et al, [20] developed different Machine Learning techniques to evaluate the prediction and diagnosis of heart disease. The main objective is to evaluate the different classification techniques such as J48, Decision Tree, KNN and Naïve Bayes. After this, evaluating some performance in measures of accuracy, precision, sensitivity, specificity are evaluated .

Data Source

Clinical databases have collected a significant amount of information about patients and their medical conditions. Records set with medical attributes were obtained from the Cleveland Heart Disease database. With the help of the dataset, the patterns significant to the heart attack diagnosis are extracted. The records were split equally into two datasets: training dataset and testing dataset. A total of 303 records with 76 medical attributes were obtained. All the attributes are numeric-valued. We are working on a reduced set of attributes, i.e. only 14 attributes.

All these restrictions were announced to shrink the digit of designs, these are as follows:

1. The features should seem on a single side of the rule.
2. The rule should distinct various features into the different groups.
3. The count of features available from the rule is organized by medical history of people having heart disease only.

SYSTEM ANAYLSIS

Existing System

Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. There are many ways that a medical misdiagnosis can present itself. Whether a doctor is at fault, or hospital staff, a misdiagnosis of a serious illness can have very extreme and harmful effects. The National Patient Safety Foundation cites that 42% of medical patients feel they have had experienced a medical error or missed diagnosis. Patient safety is sometimes negligently given the back seat for other concerns, such as the cost of medical tests, drugs, and operations. Medical Misdiagnoses are a serious risk to our healthcare profession. If they continue, then people will fear going to the hospital for treatment. We can put an end to medical misdiagnosis by informing the public and filing claims and suits against the medical practitioners at fault.

Disadvantages:

- Prediction is not possible at early stages.
- In the Existing system, practical use of collected data is time consuming.
- Any faults occurred by the doctor or hospital staff n predicting would lead to fatal incidents.
- Highly expensive and laborious process needs to be performed before treating the patient to find out if he/she has any chances to get heart disease in future.

Proposed System

This section depicts the overview of the proposed system and illustrates all of the components, techniques and tools are used for developing the entire system. To develop an intelligent and user-friendly heart disease prediction system, an efficient software tool is needed in order to train huge datasets and compare multiple machine learning algorithms. After choosing the robust algorithm with best accuracy and performance measures, it will be implemented on the development of the smart phone-based application for detecting and predicting heart disease risk level. Hardware components like Arduino/Raspberry Pi, different biomedical sensors, display monitor, buzzer etc. are needed to build the continuous patient monitoring system.

ALGORITHMS

Logistic Regression

A popular statistical technique to predict binomial outcomes (y = 0 or 1) is Logistic Regression. Logistic regression predicts categorical outcomes (binomial / multinomial values of y). The predictions of Logistic Regression (henceforth, LogR in this article) are in the form of probabilities of an event occurring, i.e. the probability of y=1, given certain values of input variables x. Thus, the results of LogR range between 0-1.

LogR models the data points using the standard logistic function, which is an S- shaped curve also called as sigmoid curve and is given by the equation:

$$\frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}}$$

Logistic Regression Assumptions:

- Logistic regression requires the dependent variable to be binary.
- For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
- Only the meaningful variables should be included.
- The independent variables should be independent of each other·
- Logistic regression requires quite large sample sizes.
- Even though, logistic (**logit**) regression is frequently used for binary variables (2 classes), it can be used for categorical dependent variables with more than 2 classes.
- In this case it’s called Multinomial Logistic Regression.

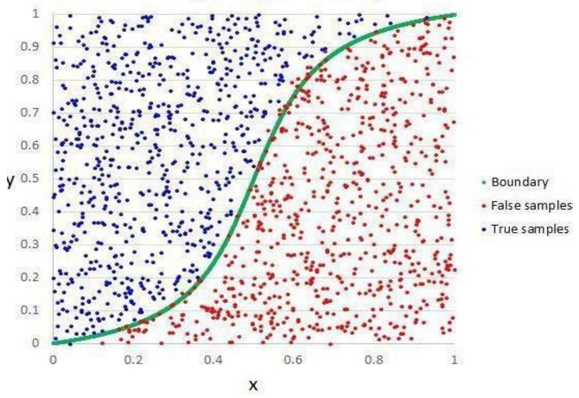


Fig 3.1: logistic regression

Random Forest

Random forest is a supervised learning algorithm which is used for both classification as well as regression .But however ,it is mainly used for classification problems .As we know that a forest is made up of trees and more trees means more robust forest .

Similarly ,random forest creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting .It is ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result .

Working of Random Forest with the help of following steps:

- First ,start with the selection of random samples from a given dataset.
- Next ,this algorithm will construct a decision tree for every sample .Then it will get the prediction result from every decision tree .
- In this step, voting will be performed for every predicted result.
- At last ,select the most voted prediction results as the final prediction result. The following diagram will illustrates its working-

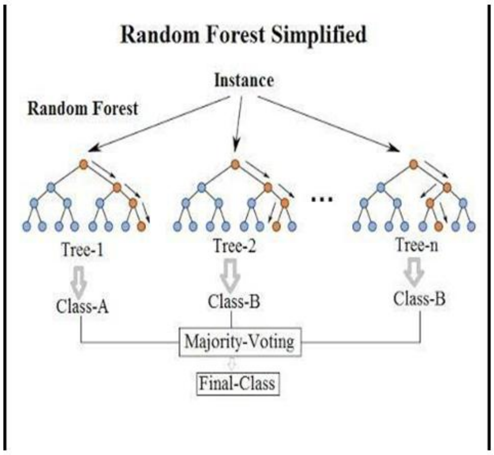


Fig 3.2: Random Forest

Feasibility Study
A Feasibility Study is a preliminary study undertaken before the real work of a project starts to ascertain the likely hood of the projects success. It is an analysis of possible alternative solutions to a problem and a recommendation on the best alternative.

SYSTEM ARCHITECTURE

The below figure shows the process flow diagram or proposed work. First we collected the Cleveland Heart Disease Database from UCI website then pre-processed the dataset and select 16 important features.

Output

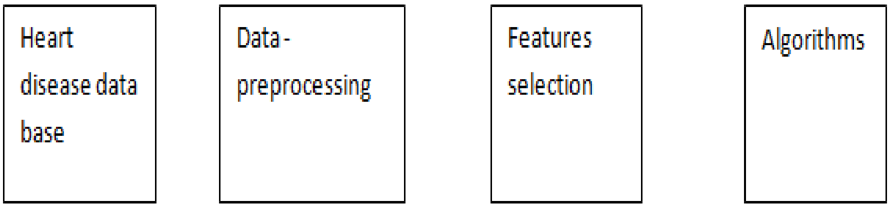


Fig 5.1: System Architecture

For feature selection we used Recursive feature Elimination Algorithm using Chi2 method and get 16 top features. After that applied ANN and Logistic algorithm individually and compute the accuracy. Finally, we used proposed Ensemble Voting method and compute best method for diagnosis of heart disease.

Modules
The entire work of this project is divided into 4 modules. They are:

a) Data Pre-processing
b) Feature
c) Classification
d) Prediction

Data Pre-processing:
This file contains all the pre-processing functions needed to process all input documents and texts. First we read the train, test and validation data files then performed some preprocessing like tokenizing, stemming etc. There are some exploratory data analysis is performed like response variable distribution and data quality checks like null or missing values etc.

Feature:
Extraction In this file we have performed feature extraction and selection methods from sci- kit learn python libraries. For feature selection, we have used methods like simple bag-of- words and n-grams and then term frequency like tf-tdf weighting. We have also used
word2vec and POS tagging to extract the features, though POS tagging and word2vec has not been used at this point in the project.

Classification:
Here we have built all the classifiers for the breast cancer diseases detection. The extracted features are fed into different classifiers. We have used Naive-bayes, Logistic Regression, Linear SVM, Stochastic gradient decent and Random forest classifiers from sklearn. Each of the extracted features was used in all of the classifiers. Once fitting the model, we compared the f1 score and checked the confusion matrix.

After fitting all the classifiers, 2 best performing models were selected as candidate models for heart diseases classification. We have performed parameter tuning by implementing Grid Search CV methods on these candidate models and chosen best performing parameters for these classifier.

Finally selected model was used for heart disease detection with the probability of truth. In Addition to this, we have also extracted the top 50 features from our term-frequency tfidf Vectorizer to see what words are most and important in each of the classes.

We have also used Precision-Recall and learning curves to see how training and test set performs when we increase the amount of data in our classifiers.

Prediction:
Our finally selected and best performing classifier was algorithm which was then saved on disk with name final_model.sav. Once you close this repository, this model will be copied to user's machine and will be used by prediction.py file to classify the Heart diseases . It takes a news article as input from user then model is used for final classification output that is shown to user along with probability of truth.

Use-Case Diagram
A use case diagram is a diagram that shows a set of use cases and actors and their relationships. A use case diagram is just a special kind of diagram and shares the same common properties as do all other diagrams, i.e a name and graphical contents that are a projection into a model. What distinguishes a use case diagram from all other kinds of diagrams is its particular content.

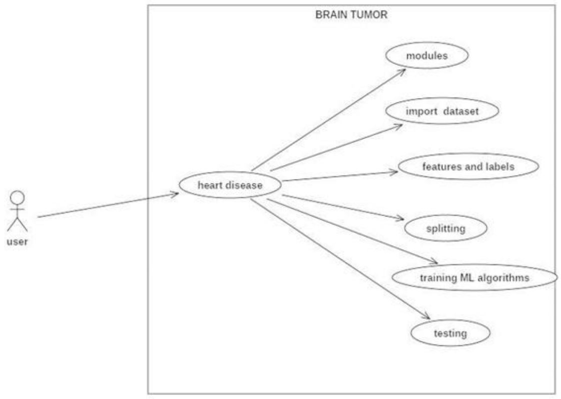


Fig 5.4: Use case Diagram

Effort, Duration and Cost Estimation Using Cocomo Model The Cocomo (Constructive Cost Model) model is the most complete and thoroughly documented model used in effort estimation. The model provides detailed formulas for determining the development time schedule, overall development effort, and effort breakdown by phase and activity as well as maintenance effort.

COCOMO estimates the effort in person months of direct labor. The primary effort factor is the number of source lines of code (SLOC) expressed in thousands of delivered source instructions (KDSI).The model is developed in three versions of different level of detail basic, intermediate, and detailed. The overall modeling process takes into account three classes of systems.

1. **Embedded:** This class of system is characterized by tight constraints, changing environment, and unfamiliar surroundings. Projects of the embedded type are model to the company and usually exhibit temporal constraints.
2. **Organic**: This category encompasses all systems that are small relative to project size and team size, and have a stable environment, familiar surroundings and relaxed interfaces. These are simple business systems, data processing systems, and small software libraries.
3. **Semidetached**: The software systems falling under this category are a mix of those of organic and embedded in nature.

Some examples of software of this class are operating systems, database management system, and inventory management systems.

For basic COCOMO Effort = a*(KLOC) b Type =c*(effort)d

For Intermediate and Detailed COCOMO Effort = a * (KLOC) b* EAF (EAF = product of cost drivers)

Table 3.1: Organic, Semidetached and Embedded system values

Type of Product	A	B	C	D
Organic	2.4	1.02	2.5	0.38
Semi Detached	3.0	1.12	2.5	0.35
Embedded	3.6	1.20	2.5	0.32

Intermediate COCOMO model is a refinement of the basic model, which comes in the function of 15 attributes of the product. For each of the attributes the user of the model has to provide a rating using the following six point scale.

VL(Very Low)
LO (Low)
NM(Nominal)

HI(High)
VH (VeryHigh)
XH (ExtraHigh)

The list of attributes is composed of several features of the software and includes product, computer, personal and project attributes as follows.

Product Attributes

- **Required reliability (RELY):** It is used to express an effect of software faults ranging from slight inconvenience (VL) to loss of life (VH). The nominal value (NM) denotes moderate recoverable losses.
- **Data bytes per DSI (DATA):** The lower rating comes with lower size of a database. Complexity (CPLX): The attribute expresses code complexity again ranging from straight batch code (VL) to real time code with multiple resources scheduling (XH)

Computer Attributes

- **Execution time (TIME) and memory (STOR) constraints:** This attributeidentifies the percentage of computer resources used by the system. NM states that less than 50% is used; 95% is indicated by XH.]
- **Virtual machine volatility (VIRT):** It is used to indicate the frequency of changes made to the hardware, operating system, and overall software environment. More frequent and significant changes are indicated by higher ratings.
- **Development turnaround time (TURN):** This is a time from when a job is submitted until output becomes received. LO indicated a highly interactive environment, VH quantifies a situation when this time is longer than 12 hours.

IMPLEMENTATION

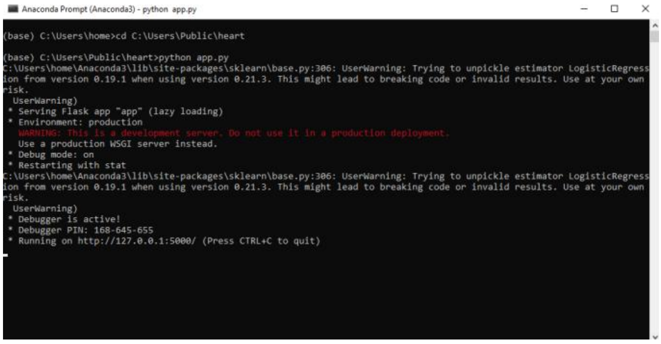
Steps for Implementation

1. Install the required packages for building the ‘Passive Aggressive Classifier’.
2. Load the libraries into the workspace from the packages.
3. Read the input data set.
4. Normalize the given input dataset.
5. Divide this normalized data into two parts:

a) Train data
b) Test data (Note: 80% of Normalized data is used as Train data, 20% of the Normalized data is used as Test data.)

Anaconda Prompt:

Screen Shots



Anaconda Prompt (Anaconda3) - python app.py

```
(base) C:\Users\home>cd C:\Users\Public\heart

(base) C:\Users\Public\heart>python app.py
C:\Users\home\Anaconda3\lib\site-packages\sklearn\base.py:306: UserWarning: Trying to unpickle estimator LogisticRegress
ion from version 0.19.1 when using version 0.21.3. This might lead to breaking code or invalid results. Use at your own
risk.
  UserWarning)
 * Serving Flask app "app" (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: on
 * Restarting with stat
C:\Users\home\Anaconda3\lib\site-packages\sklearn\base.py:306: UserWarning: Trying to unpickle estimator LogisticRegress
ion from version 0.19.1 when using version 0.21.3. This might lead to breaking code or invalid results. Use at your own
risk.
  UserWarning)
 * Debugger is active!
 * Debugger PIN: 168-645-655
 * Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
```

Fig 8.1: Anaconda Prompt

Home Screen:

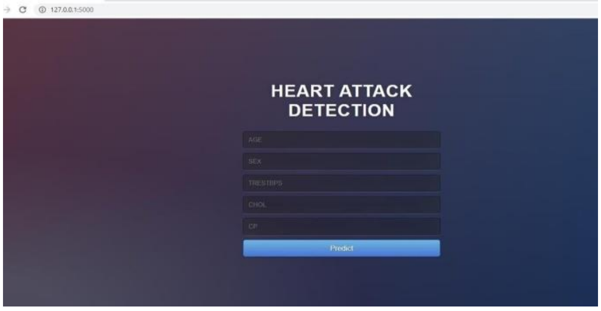
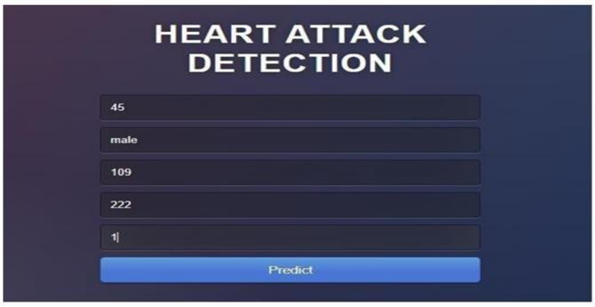


Fig 8.2: Home Screen for Heart Attack Detection

Input:



HEART ATTACK DETECTION

45

male

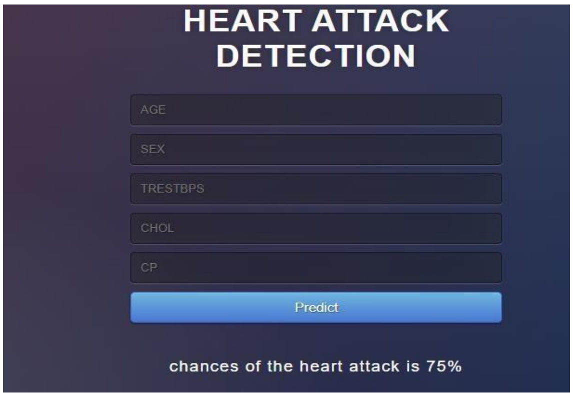
109

222

1

Fig8.3: Patient Details

Output:



HEART ATTACK DETECTION

AGE

SEX

TRESTBPS

CHOL

CP

Predict

chances of the heart attack is 75%

Fig 8.4: Output for particular patient details

CONCLUSION

In this project, we introduce about the heart disease prediction system with different classifier techniques for the prediction of heart disease. The techniques are Random Forest and Logistic Regression: we have analyzed that the Random Forest has better accuracy as compared to Logistic Regression. Our purpose is to improve the performance of the Random Forest by removing unnecessary and irrelevant attributes from the dataset and only picking those that are most informative for the classification task.

Future Scope

As illustrated before the system can be used as a clinical assistant for any clinicians. The disease prediction through the risk factors can be hosted online and hence any internet users can access the system through a web browser and understand the risk of heart disease. The proposed model can be implemented for any real time application .Using the proposed model other type of heart disease also can be determined. Different heart diseases as rheumatic heart disease, hypertensive heart disease, ischemic heart disease, cardiovascular disease and inflammatory heart disease can be identified.

Other health care systems can be formulated using this proposed model in order to identify the diseases in the early stage. The proposed model requires an efficient processor with good memory configuration to implement it in real time. The proposed model has wide area of application like grid computing, cloud computing, robotic modeling, etc. To increase the performance of our classifier in future, we will work on ensembling two algorithms called Random Forest and Adaboost. By ensembling these two algorithms we will achieve high performance.

REFERENCES

[1]. P .K. Anooj, ―Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules‖; Journal of King Saud University – Computer and Information Sciences (2012) 24, 27–40. Computer Science & Information Technology (CS & IT) 59

[2]. Nidhi Bhatla, Kiran Jyoti"An Analysis of Heart Disease Prediction using Different Data Mining Techniques". International Journal of Engineering Research & Technology

[3]. Jyoti Soni Ujma Ansari Dipesh Sharma, Sunita Soni. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction".

[4]. Chaitrali S. Dangare Sulabha S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" International Journal of Computer Applications (0975 – 888)

[5]. Dane Bertram, Amy Voida, Saul Greenberg, Robert Walker, "Communication, Collaboration, and Bugs: The Social Nature of Issue Tracking in Small, Collocated Teams".

[6]. M. Anbarasi, E. Anupriya, N. Ch. S. N. Iyengar, ―Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm‖; International Journal of Engineering Science and Technology, Vol. 2(10), 2010.

[7]. Ankita Dewan, Meghna Sharma," Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification", 2nd International Conference on Computing for Sustainable Global Development IEEE 2015 pp 704-706. [2].

[8]. R. Alizadehsani, J. Habibi, B. Bahadorian, H. Mashayekhi, A. Ghandeharioun, R. Boghrati, et al., "Diagnosis of coronary arteries stenosis using data mining," J Med Signals Sens, vol. 2, pp. 153-9, Jul 2012.

[9]. M Akhil Jabbar, BL Deekshatulu, Priti Chandra," Heart disease classification using nearest neighbor classifier

with feature subset selection", Anale. Seria Informatics, 11, 2013

[10]. Shadab Adam Pattekari and Asma Parveen," PREDICTION SYSTEM FOR HEART DISEASE USING NAIVE BAYES", International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624, Vol 3, Issue 3, 2012, pp 290-294.

[11]. C. Kalaiselvi, PhD, "Diagnosis of Heart Disease Using K-Nearest Neighbor Algorithm of Data Mining", IEEE, 2016

[12]. Keerthana T. K., " Heart Disease Prediction System using Data Mining Method", International Journal of Engineering Trends and Technology", May 2017.

[13]. Data Mining Concepts and Techniques, Jiawei Han and Micheline Kamber, EL SEVIER.

[14]. Animesh Hazra, Arkomita Mukherjee, Amit Gupta, Prediction Using Machine Learning and Data Mining July 2017, pp.2137-2159.