Web-based Application to Detect Heart Attack using Machine Learning

Rachit Singh[1], Gautam Saw[2], Rupesh Yadav[3], Yash Sawant[4]

[1,2,3,4]*B.E. Computers, Rajiv Gandhi Inst. of Technology, Mumbai, India*

Abstract - *an aphorism goes that "we are living in an information age." The healthcare industry generates a huge amount of data daily. However, most of it is not effectively used. Efficient tools to extract knowledge from databases for clinical detection of diseases or other purposes are not much prevalent. Heart Attack is one of the most imperative problems of the healthcare industry. The diagnosis of heart disease is a complicated task, where the medical experts have to carefully scrutinize the various clinical and pathological data of the patient obtained from various tests and their reports. The advent of Machine learning along with advancements in the field of information technology led us to present a paper with an aim to reduce the time and efforts put in by doctors by automating the risk prediction with the help of machine learning classification techniques. It solves the issue by building an interactive prediction system that gives the vulnerability of an individual to heart attack, measured as a risk factor. Implementation of such a system with a simplistic web based graphic user interface is presented in this paper. The results will be made available to the user along with an alert system on their mobile phones.*

Key Words: **Heart disease, Machine learning, Classification algorithms, Prediction.**

1. INTRODUCTION

A. Heart Attack

A heart attack happens when there is a sudden complete blockage of an artery that supplies blood to an area of the heart. Our heart is a muscle and to keep it healthy, it needs a good blood supply. As we grow older, the inner walls of the arteries that supply the blood to the heart can become damaged and narrow due to the build-up of fatty materials, called plaque. When the plaque breaks, the parts of blood i.e. cells and other parts accumulate at the plaque area and form blood clots. When this blood clot completely blocks the flow of blood to the heart muscle, heart attack occurs. As a result, some of the heart muscles starts to die along with severe chest pain.

The longer the blockage is left untreated, the more the heart muscle gets damaged. The blood flow must be restored quickly else the damage to the heart muscle is permanent and fatal.

The World Health Organization has stated heart attack to be one of the prime causes of death. This illness becomes particularly life threatening when not monitored.

B. Machine Learning

By the term itself, we can comprehend that Machine Learning is something that refers to the capability of the computers to grasp knowledge automatically from experiences without being programmed straightforwardly. It is a subdivision of artificial intelligence. It focuses towards giving computers the ability to learn without being directly programmed or fed instructions.

We humans use our sensory organs such as eyes, ears, or sense of touch to get data from our surroundings and then use this data in various ways to achieve certain goals. The most common goal of all is to make accurate predictions about the future. In other words, we learn. Now the question arises i.e. can computers learn from past experiences (or data), just like human beings? So the answer is yes. We can give the computers the ability to learn and predict future events - basically by providing it the data and the algorithms, without explicitly programming it.

2. LITERATURE SURVEY

This system uses multilayer perceptron architecture of neural network. In this system, input is taken as 13 attributes followed by network training which is completed further with training data by the algorithm i.e. back-propagation learning algorithm. This prediction system gives maximum accuracy of about 98.58% with same heart disease database for 20 neurons in hidden layer. With 5 neurons in the hidden layer, it gives maximum accuracy of 93.39% with running time of 3.86s. This system uses Self organizing map to develop a useful and accurate technique for retrieval and classification [1].

For the classification of cardiac arrhythmia, four classifiers were used. This includes Support Vector Machine, Logistic Regression, Random Forest Algorithm and KNN classifier. The dataset was gone for the process of cross validation and testing, the result concluded that Support Vector Machine Classifier gives the maximum accuracy, which was around 91.2%. Thus, they used Support Vector Machine Classifier for classifying arrhythmia for getting the best accurate results [2].

Alasker made a system that is used to predict Chronic Kidney Disease which is also an important factor that leads to heart disease problem. The aim of this paper was to predict Kidney disease which is done by processing and evaluating six algorithms. This is done based on the classification performance. This includes Na�e Bayes, One rule and decision table, J48, KNN, Multiplayer Perceptron. WEKA data mining tool was used to process these algorithms. The assessments of the performance of these algorithms were based on sensitivity, specificity, accuracy, model testing time, RMSE and ROC area. After performing all these algorithms, Na�e Bayes algorithm comes out as better classification algorithm in terms of accuracy and performance [3].

In this system, there are various methods for gathering the readings with the help of tools like Smart mouse (detects heart rate and temperature), Smart mirror (detects heart rate using facial imaging), Smart Chair (a built-in Electronic Stethoscope for heart rate accuracy) and the smart phone app (detects via finger taps). For the predictions to reach peak accuracy the system needs to be fed with multiple data records from the above tools. With a good size data set ready, the Neural network prediction algorithm is on mark to produce optimal results. Each data entries done will affect the results of the functioning of algorithm. The accuracy of these predictions marks up to 90% [4].

3. LITERATURE SURVEY

In this paper, a web-based system that comprises of a binary classification model to predict the risk factor of an individual is proposed. The system is well equipped with a simplistic yet prolific and understandable graphic user interface. The classification follows supervised learning. The reference to dataset was collected as unstructured data in the form of medical reports and converted to a structured dataset. There are a total of 14 attributes, out of which 13 are predictor variables while the 14th one is the response variable. In the end, a web based application will be developed which can be easily used by doctors. Also an alert system is integrated in the application which will send alerts on the mobile phones of patients i.e. the risk factor (high/low).

A. Dataset Analysis

We have taken 14 attributes which are responsible for heart attack.

1. Sex

2. Chest Pain Type

3. Fasting Blood Sugar

4. Rest ECG

5. Exang - exercise induced angina

6. Slope - the slope of the peak exercise ST segment

7. CA - major vessels coloured by fluoroscopy

8. Thalach – maximum heart rate achieved

9. Rest Blood Pressure

10. Serum Cholesterol

11. Oldpeak ST depression

12. Age in Year

14. Response Variable

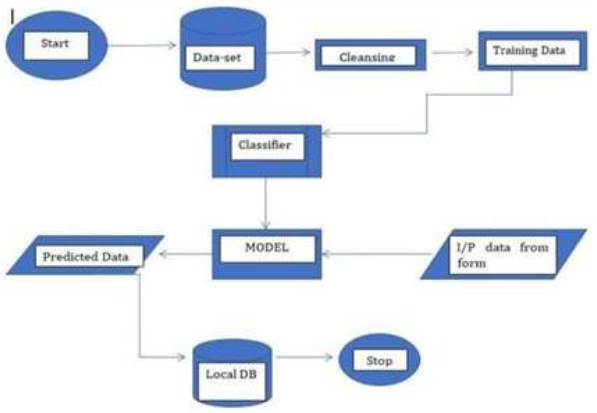


Fig -1: Flow diagram of proposed system

Fig.1 shows the flow diagram for using the web interface for obtaining the risk factor of the patient. An interactive web interface will be developed to access the classifier and check the risk factor. The backend consists of a script written in Python. It will accept the medical form data as input to the classification model and will predict the risk factor of the individual i.e. high risk or low risk.

4. IMPLEMENTATION

A. Data Cleaning and Preprocessing

Real world data is generally incomplete (missing values/attributes), noisy (containing errors or outliers) and inconsistent (inconsistent names/values) which can lead to faulty prediction. Hence data is pre-processed in order to remove the noise and inconsistency from the dataset. Data cleaning is necessary in order to fill the missing values, smooth the noisy data, identify or remove outliers and resolve inconsistencies. Scikit-learn library comprises of data preprocessing function to fill the missing values.

B. Classification Algorithms

Classification is a supervised learning method in which the computer program learns or gets trained from the dataset given as an input to it and then uses it to classify new data or observations into different categories or classes.

1) Logistic Regression – This classification algorithm finds the best apposite model to describe the relationship between the dependent variable i.e. response variable and a set of independent i.e. predictor variables. In the heart attack detection model, logistic regression is used to classify between low chances of heart attack and high chances of heart attack. Here if it predicts 0 which means it has less chances of having a heart attack and if it predicts 1 then the one has high chances of getting heart attack.

2) Na�e Bayes Classifier - It is a classification technique based on Bayes Theorem with an assumption of independence among predictors i.e. the classifier here assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For classification, Gaussian Naive Bayes algorithm was used from the scikit-learn package. Scikit-learn encapsulate multiple data mining tools. The processed dataset was spilt into training and test data. The training data was converted into a NumPy array and Gaussian Naive Bayes algorithm was applied. This regression resulted in a model to which the test data was fed in and the accuracy and other parameters were calculated. In the proposed system Naive Bayes is one of the most efficient classification algorithms.

$$p(C_k \mid x_1, \ldots, x_n) \tag{1}$$

Naive Bayes is a conditional probability model. Consider a problem instance represented by a vector x=(x1, ... x2) representing some n features (independent variables) that needs to be classified. It assigns to this instance probabilities for each of K possible outcomes.

The problem encountered with the above formulation is that if the number of features n is large then basing such a model on probability tables is infeasible. So we reformulate the model to make it more tractable. By using Bayes' theorem, the conditional probability can be stated as

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})} \tag{2}$$

3) Decision Tree Classification - Decision tree builds classification models as a tree structure. It breaks down the data set into smaller subsets and at the same time an associated decision tree is incrementally developed. As a result, we get a tree with two types of nodes - decision nodes and leaf nodes. A decision node splits up into multiple branches and a leaf node depicts a classification i.e. the classes or we can say a decision. Root node is the topmost decision node in a tree which corresponds to the best predictor.

The Decision Tree Classification Algorithm generates a decision tree and it uses the tree to classify data points into different classes.

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \tag{3}$$

Entropy is the amount of impurity in the set of features (attributes). To find the entropy of each attribute, we find the Information Gain of the corresponding attribute. Then we find the final Gain.

$$Gain(T, X) = Entropy(T) - Entropy(T, X) \tag{4}$$

The final Gain of each attribute is compared and the one with the maximum gain is chosen as the 'root node'.

4) Random Forest Classification – This classification algorithm operates by generating multiple decision trees at the time of training and outputs the class that is the mode of the classes (classification) or mean (regression) of the individual trees. Random decision forests are used because they correct the habit of over fitting of decision trees.

Random Forest Classification is the most efficient classification algorithm among all if we increase no. of trees feature. But as we increase the no of trees it shows overfitting issue which can be resolved by optimizing a tuning parameter that governs the number of features that are randomly chosen to grow each tree from the data.

C. Prediction and Visualization

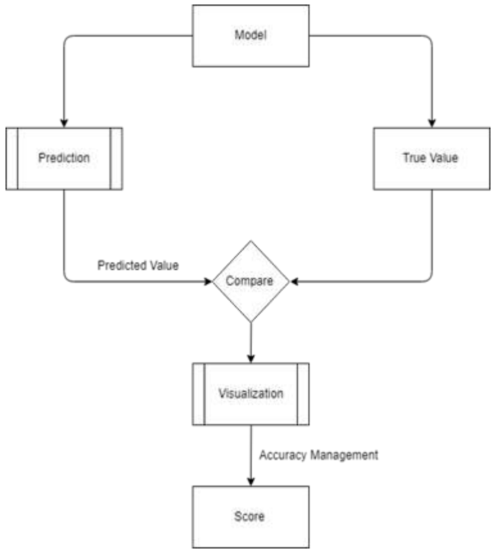


Fig -2: Prediction and visualization flow

1) Prediction - In this proposed system, first we have to fit the training data into classifier using fit() function included in scikit-learn library. Then the predict function i.e. pred() is also included in the same library is used to predict the target variables.

2) Visualization - The techniques by which the data or information is first encoded as visual objects (e.g. points, lines and bars) and then conveyed is known as data visualization. It helps in achieving our goal of communicating information clearly and efficiently to the users. The two visualization techniques sed in this system are -

1) Matplotlib - matplotlib is a plotting library which is used to create graphs and plots using Python programming language. NumPy is its numerical mathematics extension. It provides API i.e. Application programming interface which can be used to embed plots into applications using toolkits like Tkinter, wxPython, Qt, or GTK+.

2) Seaborn **-** It is built on top of matplotlib and closely integrated with pandas data structures. The plotting functions operate on whole datasets and produce informative plots by statistically aggregating and mapping the data.

5. DESIGN

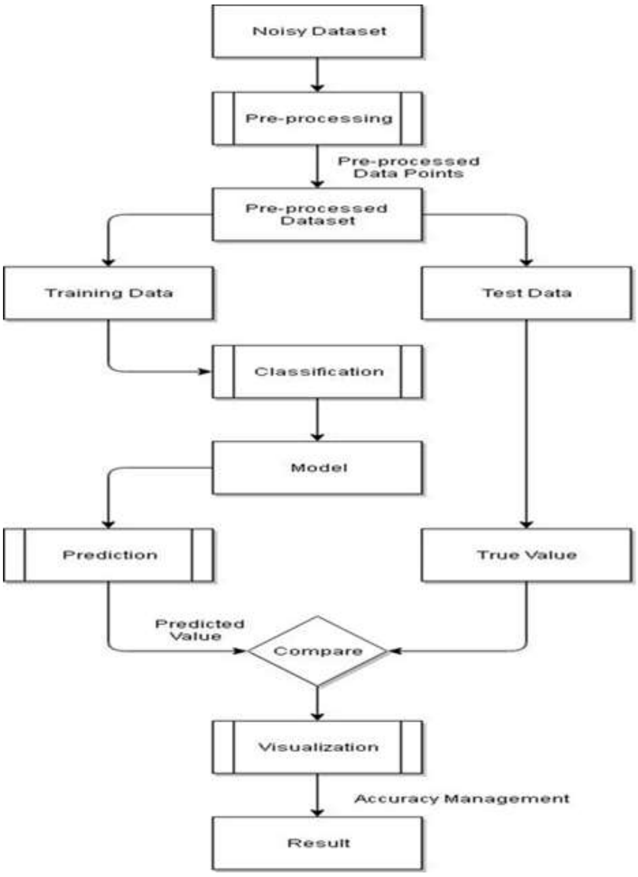


Fig -3: Model Architecture

We have the noisy data set in which there is the possibility that some of the values in the data points are missing. So, we will take that data set and pass them for pre-processing. In pre-processing, the missing values are obtained which is necessary for the further process. After pre-processing of the data set, the data is being split into two parts. In first one, the training of the data takes place based on the classification of algorithms, which is then pass into the model. From the training set we teach the model how to react, then test set is used for prediction and then the target value is predicted. Depending upon the algorithm we obtain the predicted values. Predicted value is then compared with the true values and final result is displayed. Using Visualization process we can visualize each and every attribute individually and then it is displayed with the result. The above process is repeated for every classification algorithm mentioned in the proposed system. And the best one is chosen by the most accurate prediction.

A. User Interface

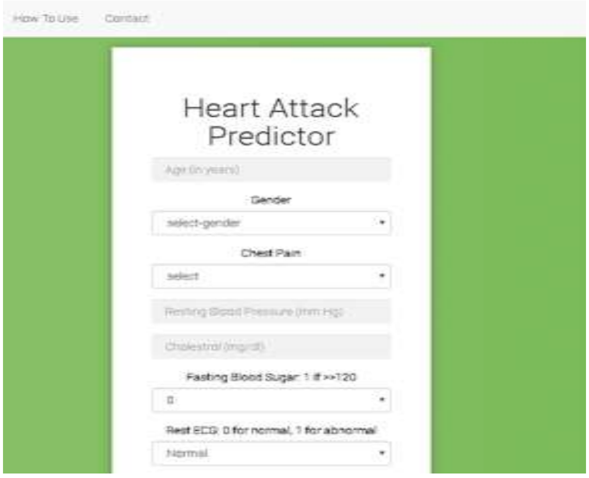


Fig -4: User Interface 1

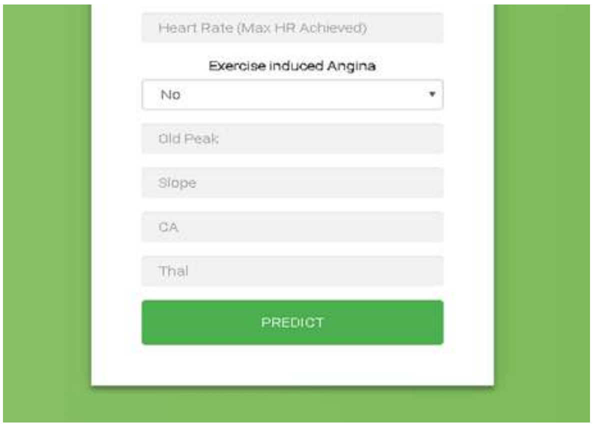
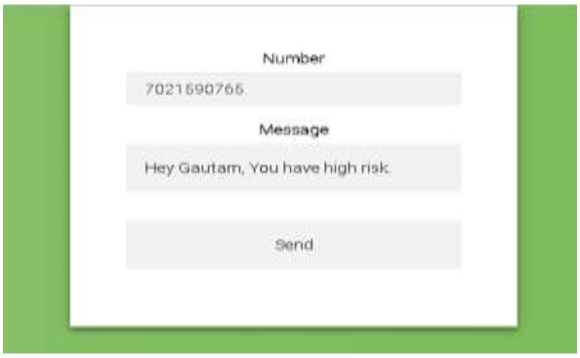


Fig -5: User Interface 2

A website is created which is based on the questionnaire where questions are health related. Basically, when a doctor inputs the values and submits, it stores into the database. Then the answers are pre-processed and stored into new dataset which is passed into the generated model. Then the prediction is done and the final score is stored into the database, from this we can get the risk factor and visualize the results and display on the website.

We have integrated an alert system in our web application where the doctor can notify the patient. The patient will be notified on his phone. An SMS will be sent via our application on the patient's phone.



Number

7021590765

Message

Hey Gautam, You have high risk.

Send

Fig -6: Alert Tab

6. RESULTS AND ANALYSIS

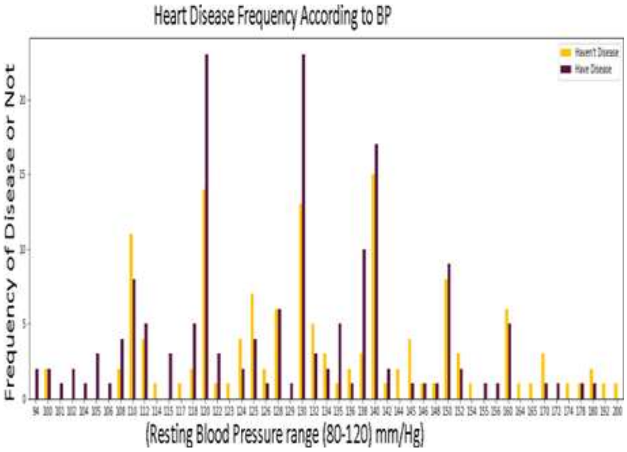


Fig -7: Heart Disease frequency for Blood Pressure

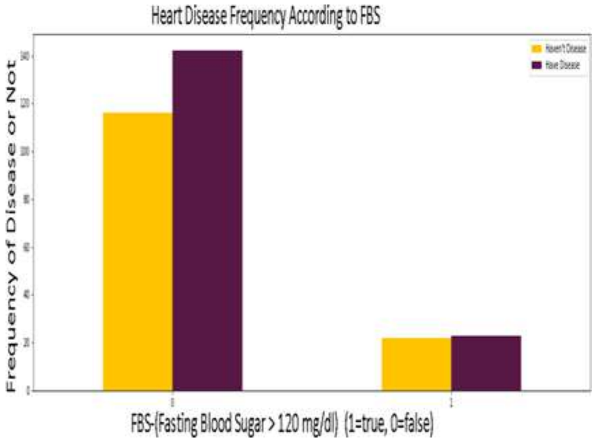


Fig -8: Heart Disease frequency for Blood Sugar

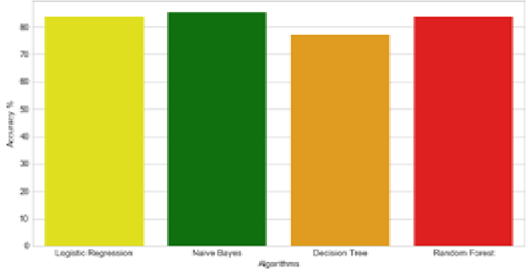
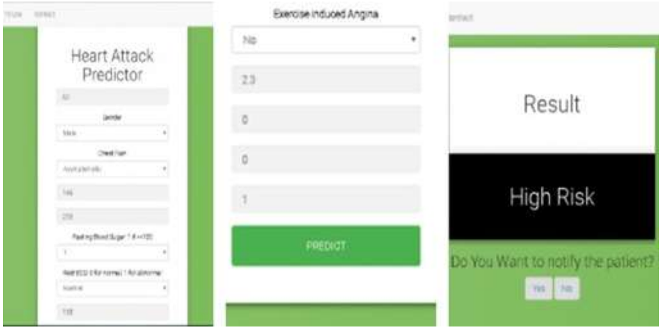


Fig -9: Comparison of various classification algorithms



Heart Attack Predictor

Gender

Male

Chest Pain

Fasting Blood Sugar

Exercise Induced Angina

No

2.3

0

0

1

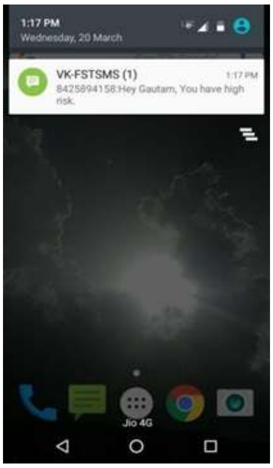
PREDICT

Result

High Risk

Do You Want to notify the patient?

Fig -10: Result predicting High Risk for an individual



1:17 PM

Wednesday, 20 March

VK-FSTSMS (1) 1:17 PM

8425894158:Hey Gautam, You have high risk.

Fig -11: Alert on patient's phone

Consider the following Confusion Matrix:

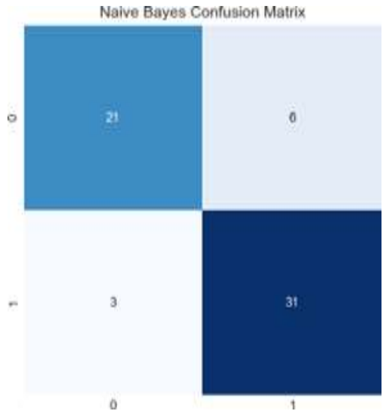


Fig -12: Confusion matrix for Na�e Bayes

Accuracy = (TP+TN)/total (5)

= 52/61 = 0.8524 = 85.24%

True Positive Rate=Recall=TP/(TP+FN) (6)

= 21/34 = 0.6176 = 61.76%

False Positive Rate=FP/(FP+TN) (7)

= 6/21 = 0.2222 = 22.22%

Specificity = TN/Actual Negatives (8)

= 21/34 = 0.6176 = 61.76%

Precision = TP/Predicted Positives (9)

= 31/37 = 0.8378 = 83.78 %

F-Score = 1/((1/recall)+(1/precision)) (10)

= 1/(34/21)+(37/31) = 0.3555

7. CONCLUSION

A prototype of the system is made to predict the chances of having a heart attack. The system calculates the risk factor if it is 0 i.e. less than 50% of blood vessels narrowing then it displays 'Low Risk' and if it is 1 i.e. more than 50% of blood vessels narrowing then it displays 'High Risk'. With this system we have also integrated an alert part where doctors can notify the patient on their phones about their condition. We have used four machine learning classifiers for the classification. Logistic regression, Naive Bayes Classifier, Decision Tree Classification and Random Forest Classification. After Validation and pre-processing of the dataset, we applied all four classifiers. Among all the classifiers, Naive Bayes Classifiers obtained maximum accuracy of 85.26%. The same classifier is implemented through web interface to calculate the risk factor.

As a part of the future scope of this system, newly introduced algorithms can be used to improve the accuracy.

REFERENCES

[1] Jayshril S. Sonawane,"Prediction of Heart Disease Using Multilayer Perceptron Neural Network"ICICES2014 - S.A.Engineering College, Chennai, Tamil Nadu, India, 2014.

[2] Prajwal Shimpi, Sanksruti Shah, "A Machine Learning Approach for the Classication of Cardiac Arrhythmia", 2017 IEEE Proceedings of the IEEE 2017 International Conference on Computing Methodologies and Communication (ICCMC), pp.6t03-607 ,2017.

[3] Haya Alasker, Lala Septem Riza, "Detection of Kidney Disease Using Various Intelligent Classifiers," 2017 3rd International Conference on Science in Information Technology (ICSITech), pp. 681-684.

[4] Rifki Wijaya, Kuspriyanto, "Preliminary Design of Estimation Heart disease by using machine learning ANN", 2013 Joint International Conference on Rural Information Communication Technology and Electric-Vehicle Technology (rICT ICeV-T) November 26-28, 2013, Bandung-Bali, Indonesia, pp.6t03-607 ,2013.

[5] Faruk Bulut, "Heart attack risk detection using Bagging classifer", 2016 IEEE 24th Signal Processing and Communication Application Conference (SIU),2016.

[6] Abdul Aziz, Aziz Ur Rehman, "Detection of Cardiac Disease using Data Mining Classication Techniques", 2017 IEEE(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No.7 ,2017.

[7] C.Sowmiyay,"Analytical Study of Heart Disease Diagnosis Using Classication Techniques" 2017 IEEE INTERNATIONAL CONFERENCE ON INTELLIGENT TECHNIQUES IN CONTROL, OPTIMIZATION AND SIGNAL PROCESSING, 2017.

BIOGRAPHIES

Mr. Rachit Singh
B.E. Computer Engineering
Interested in MIS and Machine Learning.

Mr. Gautam Saw
B.E. Computer Engineering
Interested in Software Dev.

Mr. Rupesh Yadav
B.E. Computer Engineering
Interested in Machine Learning and AI.

Mr. Yash Sawant
B.E. Computer Engineering
Interested in Machine Learning and AI.