

Heart Disease Prediction Using Machine Learning

Mansi Rajendra Gade¹ and Rasika Patil²

Department of Computer Application^{1,2}

Bharati Vidyapeeth College, Navi Mumbai, Maharashtra, India

Abstract: *Healthcare field features has immense quantity of information, for process those information bound techniques are used. Data processing is one in every of the techniques typically used. Cardiovascular disease is the major reason behind death world-wide. This technique predicts the arising prospects of Heart-Disease. However, it remains tough for clinicians to predict heart disease because it could be a complicated and expensive task. Hence, we tend to projected a medical web for predicting cardiovascular disease to assist clinicians with diagnostic and build higher selections. The end result of this technique provides whether or not the user features a heart disease or doesn't have a cardiovascular disease. The datasets are classified in terms of medical parameters. The aim of this project is to predict heart disease using data processing techniques and machine learning algorithms. This project implements five classification models scikit-learn: Logistic Regression, Support Vector Classifier, k-Nearest Neighbours, Neural Network and Random Forest Model to analyse their performance on heart information sets obtained from the UCI information repository and from Kaggle.com. The framework that may be accustomed build the project is Django.*

Keywords: Heart sickness, Cardiovascular disease, cardiac arrest, clinical Random Forest, machine learning.

I. INTRODUCTION

According to the world health organization (WHO), cardiovascular diseases are unit the primary reason behind death worldwide, with over 17.90 million individuals died in 2016[1]. CVDs are unit a group of syndromes affecting blood vessels and heart; which include heart disease (HD), that is usually expressed as coronary cardiovascular disease [2]. However, cardiovascular disease is prevented by perceive a healthy mode and avoiding risk factors. Thus, understanding what's conducive to those ugly factors might facilitate for the interference and prediction of HD. Typical, Angiography is that the primary diagnosing method; it used to confirm the localization of heart vessels' stenosis. Being expensive, long, and invasive had impelled researchers to develop automatic systems supported information gathered through a group of medical data, like data from past treatment outcomes in addition because the latest medical analysis results and databases [3]. Nowadays, machine learning techniques are used to assist clinicians in creating additional correct predictions of HD based on medical data , these data is demographic, symptom and examination, ECG, and laboratory. Many studies were carried on diagnosis and predicting heart disease using machine learning techniques [4]. Most researches have used the UCI heart disease data set due to its convenience [5]. This data set contains four sub data set and 76 to 80 class attributes; Generally, the studies that used several attributes have applied feature selection to improve relevance [6, 7]. Hence, most studies perform 14 attributes, as well as (Age, Gender, Chest pain, force per unit area ...) That are relevant for the chance factors of HD designation values [8-11]. Numerous prediction models were built using well-known machine learning techniques. The author [8] recommended a predictive model using C4.5 associate decision tree algorithms applied on the four collected and separated UCI data sets; this model achieved an accuracy of 76.06% and 75.48% for C4.5 and quick decision tree severally using only Cleveland data set. The author [7] Combined Infinite Latent feature selection method with SVM classifier and achieved an accuracy of 89.93% using three data sets, including Cleveland, Hungarian, and Switzerland, with 58 attributes. The author [6] predicted HD using the meta algorithm Adaboost on Cleveland data set and suggested reducing the number of attributes from 76 to 28 to provide higher accuracy of 80.14%. The author [12] used Alizadeh Sani data set to develop a hybrid method by enhancing the performance of Neural network using Genetic algorithm and yielded an accuracy of 93%. A comparative study using four different classifiers including SVM, KNN, C5.0, and Neural network, was approved by the author [9], he achieved a high accuracy of 93.02% by C5.0 algorithm using 14 of attributes with different data sets. Despite a substantial research output, no gold-standard model is available to predict HD. This paper aims to build a clinical decision support system allowing predicting the risk level of HD using UCI Cleveland data set. A classification model is proposed to detect

patterns in existing HD patient's data. In the next section, our methodology is described with a brief detail of the data set used. Section 3 presents the experiments and the different representations of outcomes.

II. LITERATURE REVIEW

Different research-based works have been done in the present years to find out the most preferable technique regarding heart disease prediction. In paper [13] the author has used classification algorithm Logistic Regression to build the model for Heart Disease Prediction. It is a simple prediction system. Authors have considered multiple major risk factors that are the cause of heart disease. The major risk factors are age, diabetes, hypertension, high cholesterol, tobacco smoking, alcohol intake, etc. In paper [14] the authors have used Naïve Bayes for the classification and AES(Advanced Encryption Standard) technique to provide security to user's data. The system continuously monitors the coronary heart patient and updates the data to the object converse data base and if any abnormalities are observed. In paper [15] The project involved analysis of the heart disease patient dataset with proper data processing. Then, three models were trained and tested with maximum scores Support Vector Classifier: 84.0 %, Neural Network: 83.5 %, Random Forest Classifier: 80.0 %. This project provides the deep vision into machine learning techniques for classification of heart diseases. In paper [16] authors have considered seven attributes such as Age, Sex, thal, Resting Bp, Cholesterol, Fasting Blood Sugar, Electrocardiographic which are extracted from a medical report to predict heart disease of the patient. The data mining technique used in this research paper is Naïve Bayes. In paper [17] the algorithms has been used the by authors to implement are Neural Networks (NN), KNN and SVM, to test the dataset which contains information of Algerian patients. In the end it was noticed that Neural Network gives the best results with 93% of accuracy. Paper [18] proposes a scalable system for heart disease monitoring using on Spark and Cassandra frameworks. This project is about applying real time classification for heart disease prediction system.

In [19]aim of the authors is to build an application of heart disease prediction system using robust Machine Learning algorithm which is Random Forest algorithm. A CSV file is given as input. After the successful completion of operation the result is predicted and displayed. Paper [20] uses five machine learning models namely Support Vector Machine, Random Forest, KNN, Gaussian Naive Bayes, Xg- Boost algorithms for predicting of heart disease. After performing all the classification techniques, accuracy of random forest is with 88.52% which is good and higher when compared to other models. In paper [21] authors have proposed a clinical support system for predicting heart disease to help clinicians with diagnostic and make better decisions. Machine learning algorithms such as Naive Bayes, K-Nearest Neighbour, Support Vector Machine, Random Forest, and Decision Tree are applied in this study for predicting Heart-Disease using risk factors data retrieved from medical files. The outcome reveals that Naive Bayes outperforms using both cross-validation and train-test split techniques with an accuracy of 82.17% and 84.28%, respectively. The second conclusion is that the accuracy of all algorithm decreases after applying the cross-validation technique. In paper [22] Principal Component Analysis, Hybrid Genetic Algorithm with k-Means two different kinds of data mining techniques are used for the early prediction of heart disease. The author's method reduces the dimensionality of the dataset using PCA and combined the unsupervised heuristic k-means algorithm with metaheuristic Genetic Algorithms for better combinatorial optimization. After converging, the proposed algorithm has improved the final clustering quality. The outcome reveals that these data mining techniques can predict heart disease early with an accuracy of 94.06%.

III. RESEARCH GAPS IDENTIFIED

Prediction of a specific Heart-Disease type.

The current systems that are available do not predict the Heart Disease Type such as Heart Attack, Cardio Vascular Disease, Coronary artery disease, etc. For someone who is predicted to be suffering from heart disease.

Security for User's data.

From the above literature survey what I have understood is that are very few systems that care about user's privacy and provide security to the data.

Online doctor consultation with the nearest doctor available.

When it is predicted that someone is suffering from a heart disease by the system there is no facility for the user to book an appointment with a near by doctor or to consult the a doctor online.

IV. PROPOSED METHODOLOGY

Heart disease refers to any condition affecting the heart. There are many types, some of which are preventable. Heart-Disease Prediction Using Machine Learning is a web application built on Python, Django, and Machine Learning. The web application uses following models:

1. Support Vector Classifier(SVC)
2. K-nearest neighbour(KNN)
3. Random Forest(RF)
4. Neural Network(NN)
5. Logistic Regression(LR)

The detailed architectural diagram of the system is given (Fig 1).

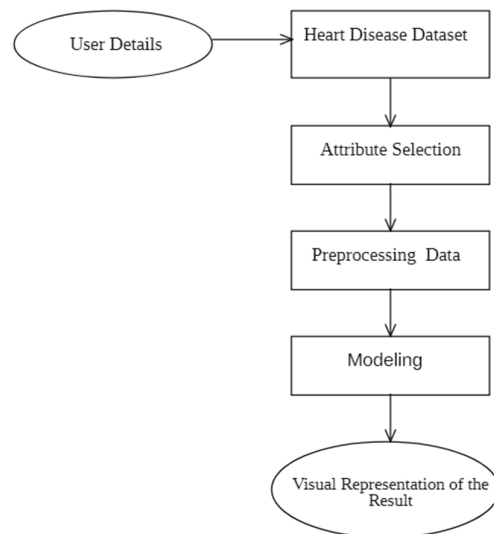


Fig 1. Architectural diagram for prediction of heart disease

User Details

In this phase User's details such as Age, Gender, Chest-Pain, Resting BP, Serum Cholesterol, Fasting Blood Sugar, Resting ECG, Maximum Heart Rate, Exercise Induced Angina, ST Depression, Slope of ST Segment, Number of Vessels, and Thal will be collected.

Heart Disease Dataset

The dataset that will be used for creating this system are available on Kaggle or UCI Repository.

Kaggle Dataset: This dataset has 16 columns and 921 rows.

UCI Heart Disease Data | Kaggle

UCI Repository Dataset: This dataset has 14 columns and 304 rows in total.

UCI Machine Learning Repository: Heart Disease Data Set

Attribute Selection

The in this phase attributes that are required would be selected using Feature Engineering. The attributes that are required for prediction are explained below in detail.

1. Age - age in years
2. Sex - (1 = male; 0 = female)
3. Cp - chest pain type
 - 0: Typical angina: chest pain related decrease blood supply to the heart
 - 1: Atypical angina: chest pain not related to heart
 - 2: Non-anginal pain: typically oesophageal spasms (non heart related)

- 3: Asymptomatic: chest pain not showing signs of disease
 - 4. Trestbps - resting blood pressure (in mm Hg on admission to the hospital) anything above 130-140 is typically cause for concern
 - 5. Chol - serum cholestorol in mg/dl
 - Serum = LDL + HDL + .2 * triglycerides
 - Above 200 is cause for concern
 - 6. Fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
'>126' mg/dl signals diabetes
 - 7. Restecg - resting electrocardiographic results
 - 0: Nothing to note
 - 1: ST-T Wave abnormality can range from mild symptoms to severe problems signals non-normal heart beat
 - 2: Possible or definite left ventricular hypertrophy enlarged heart's main pumping chamber
 - 8. Thalach - maximum heart rate achieved
 - 9. Exang - exercise induced angina (1 = yes; 0 = no)
 - 10. Oldpeak - ST depression induced by exercise relative to rest looks at stress of heart during excercise unhealthy heart will stress more
 - 11. Slope - the slope of the peak exercise ST segment
 - 0: Upsloping: better heart rate with excercise (uncommon)
 - 1: Flatsloping: minimal change (typical healthy heart)
 - 2: Downsloping: signs of unhealthy heart
 - 12. Ca - number of major vessels (0-3) colored by flourosopy
 - Colored vessel means the doctor can see the blood passing through
 - The more blood movement the better (no clots)
 - 13. Thal - thalium stress result
 - 1,3: normal
 - 6: fixed defect: used to be defect but ok now
 - 7: reversable defect: no proper blood movement when excercising
- Target - have disease or not (1=yes, 0=no) (= the predicted attribute)

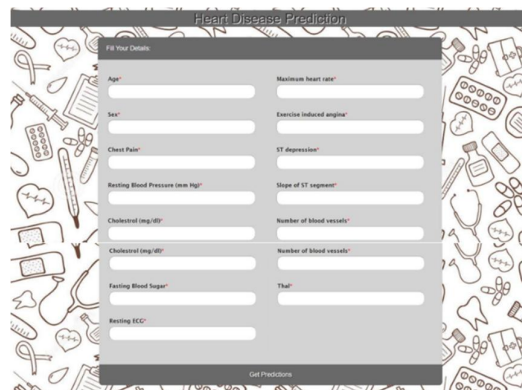


Fig 2. Home page of the system

V. DATA PRE-PROCESSING

Data preparation is the most critical first step in any predictive model; it helps to transform data into an understandable format to enhance model efficiency. Medical data are generally incomplete, lacking attribute values, and noisy since containing outliers or irrelevant data [13]. The UCI Cleveland data set used in this study contains six missing values, including four missing values for the number of major vessels (Ca) attribute and two missing values for Heart rate (Thal) attribute. The handle these missing values, we used the “Mode” imputation method that replaced missing values by the

most frequently occurring value since all missing values are categorical [14]. The predicted attribute (num) of the original data set contained 5 values; a value 0 indicated the absence of HD and values between 1 and 4 reported different levels of HD, respectively. In this study, we have tendency to have an interest within the presence or absence of HD without interest in the exact disease classification. Hence, the class attribute is reclassified into a binary value of 0 or 1, indicating the absence or presence of HD in the patients, respectively. A detailed diagram for Data pre-processing is given below.

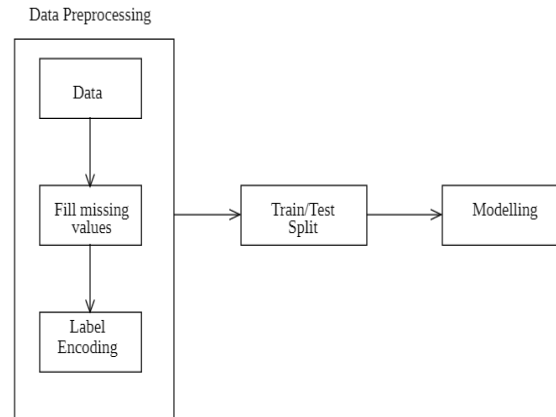


Fig 3. Pre-processing on Dataset

1. **Filling Missing Values:** In this step rows that contain missing values will be filled using mean, median method. After filling the missing values the data will then pass to the next step that is label encoding.
2. **Label Encoding:** Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.
3. **Train-Test Split:** The train test split technique can be used for classification and regression problems to test machine learning algorithms. The procedure takes the given dataset and splits it into two subsets:
 - Training dataset: it is used to train the algorithm and fit the machine learning model.
 - Test dataset: Using the input element from the training data, the algorithms make predictions.

The model is first to fit on the available data with known inputs and outputs. It is then run to make predictions on the rest of the data subset to learn from it. This can be used to make predictions on future data sets where the expected input and output values are non-existent.

5.1 Modelling

Predictive modelling is the subpart of data analytics that uses data mining and probability to predict results. Each model is built up by the number of predictors that are highly favourable to determine future decisions. Once the data is received for a specific predictor, an analytical model is formulated. The algorithms that will be used to build the model are explained in detail below.

- i. **Support Vector Classifier:** Support Vector Classifier implements Support Vector Machine based on lib SVM library. It supports different types of kernels. It is used for binary classification.
- ii. **K-Nearest Neighbor:** k-nearest neighbor also known as KNN is a supervised machine learning algorithm. It can be used for both classification and regression problems but majorly it is used for classification problems.
- iii. **Logistic Regression:** Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The target variable is binary in nature having data coded as either 1 (success/yes) or 0 (failure/no). Types of logistic regression are binary, multinomial and ordinal.
- iv. **Neural Network:** **Neural networks (NN)** are a class of machine learning (ML) algorithms. Neural networks are not algorithms, but rather frameworks for many different machine learning algorithms that work together. The algorithms process complex data. That have demonstrated good accuracy on image classification, object detection, speech recognition and natural language processing applications.

- v. Random Forest (RF), currently, is one of the most potent and popular classifications and regression algorithms used in machine learning. It has been widely used to examine data and make decisions by many researchers as classifiers in the healthcare domain. It is capable of handling large datasets with high dimensionality. Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

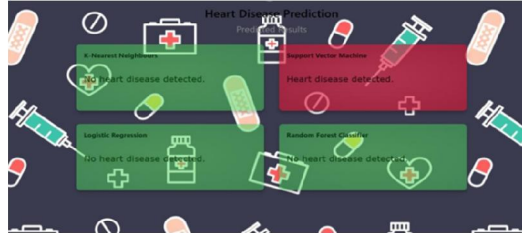


Fig 4. Heart disease predicted

Fig 4. These algorithms show that the person suffers from Heart Disease, which means the patient must visit a hospital as quickly as possible.

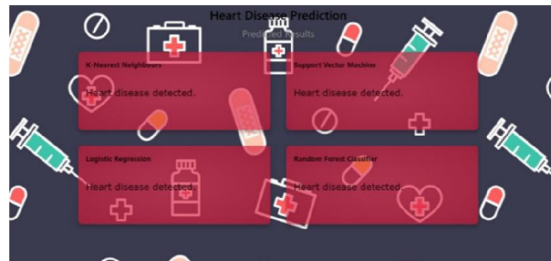


Fig 5. NO heart disease predicted

Fig 5. Three algorithms have shown that the patient does not suffer from heart disease.

Performance measurement In order to evaluate the validity of the predictive model, various measurements can be calculated suchlike sensitivity, specificity, accuracy, and precision, by using the confusion matrix (Table I). Specificity measures the proportion of negatives which are correctly identified, and sensitivity measures the percentage of real positives that are correctly identified [21]. These measures can be mathematically represented by the following formulas. Where TP, TN, FP, and FN signify True Positive (number of positive data that were correctly labelled by the classifier), True Negative (number of negative data that were correctly labelled by the classifier), False Positive (number of negative data that were incorrectly labelled as positive), and False Negative (number of positive data that were mislabelled as negative), respectively.

Table 1: Confusion Matrix

		Actual values	
		positive	negative
Predicted values	positive	TP	FP
	negative	FN	TN

$$Specificity = \frac{TN}{TN + FP} \quad (1)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

$$Precision = \frac{TP}{FP + TP} \quad (3)$$

$$Accuracy = \frac{(TN + TP)}{(TN + FP + FN + TP)} \quad (4)$$

VI. RESULT

Different experiments are conducted to evaluate the performance and also the validation of the developed model using the Cleveland trained data set with 14 totally different attributes. Results are estimated using confusion matrix measurements and to match the accuracy using totally different algorithms. For the first experiment, we used the train-test split validation technique where our data-set is divided into two parts, and we made several tests with different percentages, the best splitting we achieved is the 70% of the data for training and 30% for testing. Fig 5 Shows the results obtained by applying LR, KNN, SVM, RF, and NN algorithms. Based on the experimental results, it is clear that the classification accuracy of the KNN algorithm is the highest, followed by SVM compared to other algorithms. However, the train-test split validation technique usually causes over fitting since the evaluation may depend mainly on which data is used in the training set and which is used in the test set. Hence, the evaluation may be significantly different depending on how the split is made. Thus, we proposed the cross-validation technique to handle this problem. The cross validation is a robust preventative measure against over fitting it uses the initial training data to generate multiple mini train test splits to tune the model. In our second experiment we used 10-fold cross-validation, where the original dataset spited into 10 same size subsamples, and the accuracy is averaged over all 10 trials to get the total effectiveness of our model. As reflected, each data point gets to be in a test set once and set k1 times in the training. This significantly reduces bias and variance, since we used most of the data for fitting and in the test set. Fig 6 gives the accuracy obtained using a 10-fold cross validation technique. It can be observed from the Fig 6 that the KNN still worked better, building the model with an accuracy of 91.80%, and SVM was second with an accuracy of 86.88%. We can conclude that KNN performs better and gets better accuracy compared to other algorithms. Also, all the results' accuracy is decreased after using the cross-validation technique.

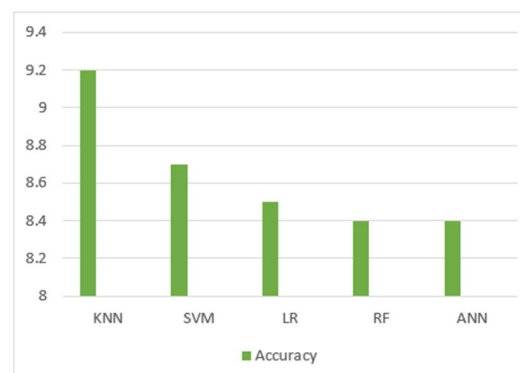


Fig. 6 Comparison of the accuracy of ML algorithms using train/test split technique.

VII. CONCLUSION

This study has been conducted to help clinicians to produce an accurate and efficient predictive system, the model validation is conducted with both cross-validation and the train-test split of data. The results showed that KNN achieved the highest accuracy compared to other algorithms using both validation techniques. Despite the accuracy is decreased once we applied the cross-validation, we tend to believe that this method is that the best in our model since the used data-set isn't large therefore the method did not take an extended time, at the same time we solved the problem of over fitting. The results were significant, and we believe that the achieved results using our predictive model based on ML algorithms could improve the knowledge on the prediction of heart disease risk through better diagnosis and interpretation thus, appropriate clinical decisions.

REFERENCES

- [1]. URL:[http://who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](http://who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))
- [2]. URL: <http://nhlbi.nih.gov>. National heart, lung, and blood institute.
- [3]. N. Mishra and S. Silakari "Predictive Analytics: A Survey, Trends, Application, Opportunities and Challenges," International Journal of computer science and information technologies, vol 3(3), pp. 4434-4438, 2012.

- [4]. H. Alharti. "Healthcare predictive analytics: An overview with a focus on Saudi Arabia," Journal of Infection and Public Health, vol 11(6), pp. 749-756, 2018.
- [5]. R. Detran Heart Disease Dataset. "Retrieved from: <http://archive.ics.edu/ml/machine-learning-databases/heartdisease/cleveland.data>" 1988.
- [6]. K. H., Miao, J. H. Miao & G. Miao. "Diagnosing Coronary Heart Disease Using Ensemble Machine Learning," International Journal of Advanced Computer Science and Applications, vol 7(10), 2016.
- [7]. L. M. Hung, D. T. Toan, & V. T. Lang. "Automatic Heart Disease Prediction Using Feature Selection and Data Mining Technique,". Journal of Computer Science and Cybernetics, vol 34(1), pp. 33-47, 2018.
- [8]. R. El-Bialy, M. A. Salamay, O. H. Karam, & M. E. Khalifa. "Feature Analysis of Coronary Artery Heart Disease Data Sets". International Conference on Communication, Management and Information Technology. Procedia Computer Science, vol 65, pp. 459-468, 2015.
- [9]. M. Abdar, R. Sharareh, N. Kalhori, T. Sutikno, I. M. I. Subroto & G. Arji. "Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases," International Journal of Electrical and Computer Engineering, vol 5(6), pp. 1569-1576, 2015.
- [10]. A. K. Paul, P. C. Shill, R. I. Rabin, & M. A. H. Akhand. "Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease,". IEEE 5th International Conference on Informatics, Electronics and Vision. 2016.
- [11]. Purushottam, K. Saxena, & R. Sharma (2016). Efficient Heart Disease Prediction System. Procedia Computer Science, 85, 962-969.
- [12]. Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, & A. A. Yarifard. "Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm," Computer Methods and Programs in Biomedicine, vol 41, pp. 19-26, 2017.
- [13]. Montu Saw, Tarun Saxena, Sanjana Kaithwas, Rahul Yadav, Nidhi Lal, "Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning", 2020 International Conference on Computer Communication and Informatics (ICCCI -2020), Pages:6, Jan. 22-24, 2020, Coimbatore, INDIA.
- [14]. Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, "Design And Implementing Heart Disease Prediction Using Naives Bayesian", Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019), Pages:06, 2019.
- [15]. Baban. U. Rindhe, Nikita Ahire, Rupali Patil, Shweta Gagare, Manisha Darade, "Heart Disease Prediction Using Machine Learning", International Journal of Advanced Research in Science, Communication and Technology (IJAR SCT), Pages:08, May 2021.
- [16]. Ninad Marathe, Sushopti Gawade, Adarsh Kanekar, "Prediction of Heart Disease and Diabetes Using Naive Bayes Algorithm", International Journal of Scientific Research in Computer Science, Engineering and Information Technology IJSRCSEIT, May-June-2021.
- [17]. Dhaiddine Salhi, Abdelkamel Tari, and M-Tahar Kechadi, "Using Machine Learning for Heart Disease Prediction", February 2021
- [18]. Abderrahmane Ed-daoudy, Khalil Maalmi, "Real-time machine learning for early detection of heart disease using big data approach", 2019, IEEE
- [19]. M. Snehith Raja, M. Anurag, Ch. Prachetan Reddy, Nageswara Rao Sirisala, "MACHINE LEARNING BASED HEART DISEASE PREDICTION SYSTEM", International Conference on Computer Communication and Informatics (ICCCI 2021), Jan 27-29, 2021
- [20]. Shaik Farzana, Duggineni Veeraiah, "Dynamic Heart Disease Prediction using Multi-Machine Learning Techniques", 2020 IEEE
- [21]. Halima EL HAMDAOUI, Saïd BOUJRAF, Nour El Houda CHAOUI, Mustapha MAAROUFI "A Clinical support system for Prediction of Heart Disease using Machine Learning Techniques", 5th International Conference on Advanced Technologies For Signal and Image Processing, 2020
- [22]. Md. Touhidul Islam, Sanjida Reza Rafa, Md. Golam Kibria "Early Prediction of Heart Disease Using PCA and Hybrid Genetic Algorithm with k-Means" International Conference on Computer and Information Technology (ICCIT), 19-21 December, 2020