Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Engineering

Elixir Comp. Engg. 165 (2022) 56200-56204

ISSN: 2229-712X

Heart Disease Prediction Analysis Using Machine LearningAlgorithms

Surya Chauhan[1] and Sandeep Rana[2]

[1]Scholar Computer Science, J.P. Institute of Engineering and Technology, Meerut.

[2]Department of Computer Science, J.P. Institute of Engineering and Technology, Meerut.

ARTICLE INFO

Article history:

Received: 20 March 2022;

Received in revised form:

20 April 2022;

Accepted: 30 April 2022;

Keywords

Data Mining,

Machine Learning,

Gaussian Naive Bayes,

Support Vector Machine (SVM),

Random Forest,

KNN(K-Nearest Neighbour),

XG-Boost.

ABSTRACT

The health care field contain large amount of dataand in order to process this data, we must use any advanced techniques that will be helpful to deliver effective results and make effective decisions on the data and obtain relevant results. Heart disease is a major problem and is one of the main reasons forsaying no. of deaths occurring worldwide. In this paper, the practical framework of Heart Disease Prediction is applied using algorithms in Machine Learning such as Logistic regression, Naïve Bayes,Support vector machine, KNN, decision tree, random forest, XG-Boost and the neural network. This framework uses 13 factors such as age, gender, blood pressure, cholesterol, oldpeak, cp, etc. In the first step, we upload a database file and selectan algorithm to perform on the selected database. Then accuracy is predicted for each selected algorithm and graph, and the model is designed for the one with the highest accuracy by training the database in it. In the next stage, input is given to each candidate parameter and based on that method produced, the stage with heart disease is predicted. We then take precautionary measures by looking at the patient's condition. Our strategy is effective in predicting the heart attack of a traumatized person. The Heart Disease Prediction Framework developed in this concept is one of the different methods that can be used within the heart disease category.

I. Introduction

Heart disease affects heart function. The World Health Organization has conducted a study and concluded that 10 million people suffer from heart disease and loss of life. The problem the healthcare industry faces in modern life is the prediction of diseases soon after a person is affected. Medical records or data are very large and data in the real world may be incomplete and inconsistent. In the past successful disease prognosis and treatment in patients may not be possible in every patient in the early stages under these conditions [1].

The burden of heart disease is growing rapidly around the world over the past few years. Numerous studies have been conducted in an effort to identify the most influential features of heart disease and to accurately predict all risks. Heart disease has even been described as a silent killer that results in the death of a person without obvious symptoms. Early diagnosis of heart disease plays an important role in making decisions about lifestyle changes inhigh-risk patients and reduces complications. This project aims to predict future Heart Disease by analyzing patient data that differentiates whether they have heart disease or not using machine learning algorithms.

Machine learning is given priority in modern life in many programs and in the field of health care. Predictability is one of the areas where machine learning plays an important role, this model is to predict heart disease by analyzing patient data i.e., the user we need to predict the risk of heart disease. Data analysis has proven its importance in the medical field. It is the basis for all of you to make any difficult decisions. Especially this analysis helps to keep a person's inclination away from the medical conclusion with the help of a balanced and appropriate treatment. By using different data mining techniques, we can test a large amount of information. In a comprehensive program of health science applications, Data Mining tends to grow. Significant efficiency and quality management of low-cost human services can be achieved through the use of data mining classification and prediction frameworks. The vast amount of information generated by medical enterprises containing encrypted data is helpful in solving powerful decisions to provide accurate results for effective decision-making. These data mining methods are used to improve the information and conclusion provided.

II. Literature Review

The literature Review summary can be listed in table 1. Various machine learning approaches are used on this popular dataset and the accuracy calculated by all the techniques is more with time computations.

III. Data Set Description

In this paper the feature vector we are considering are age, gender, chol, obesity, cp, trestbps, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal. We have taken the dataset from UCI that containing 303 records to predict the heart disease. The description of dataset feature vectors is shown in table 2.

IV. Research Methodology

Predicting Heart Disease is nothing, but it describes the condition of the heart. Heart disease is a major cause of death for both men and women. It is one of the major problems in our world. Therefore, its prediction has become a topic in one of the areas in information analysis. When the amount of data is large, it becomes difficult for the healthcare industry. Data

Tele:

E-mail address: sandeepmietcs@gmail.com

Mining and Machine Learning receives a large amount of data and converts it into useful information for making predictions and making decisions in sequence.

The primary goal is to develop a heart prediction system using various machine learning algorithms. CSV file is provided as imported. After successfully completing the task the result is predicted and displayed.

In predicting heart disease many situations rely on bizarre information because of complex problems. The different algorithms we use have different accuracy. Because of some of the complex problems, there is a growing interest in researchers to predict disease, especially in relation to the heart. In each case, they have seized it, despite obstacles we can scarcely imagine. "In this paper, heart disease predicts a structured structure that contributes to advanced research in predicting heart conditions based on clinical data provided to patients. Our approach has three phases as stated later in this study. Predictability accuracy is close to 88%.The process of research methodology is shown in figure 1.

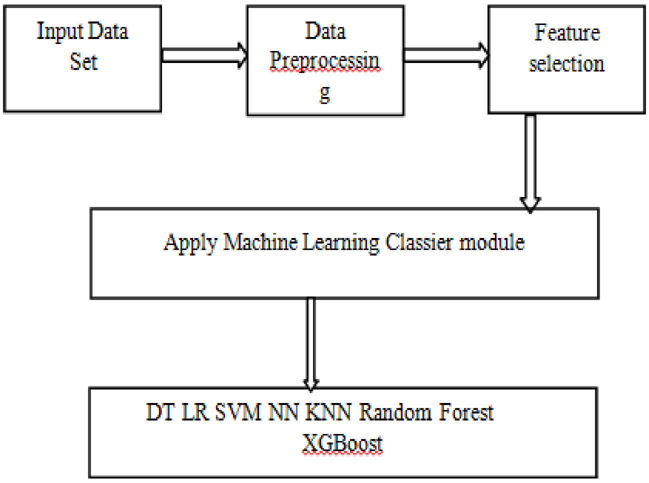

Input Data Set
Data Preprocessin g
Feature selection
Apply Machine Learning Classier module
DT LR SVM NN KNN Random Forest XGBoost

Figure 1. Research Methodology

1. Step 1 : Data Preprocessing

In the first phase of our process, the data extracted from UCI is considered an input containing 304 records. It is loaded for pre-processing. There are checkboxes available containing the algorithm models we have used in this process. The algorithms we have considered are Gaussian Naïve Bayes, Vector Support Machine, Random Forest, KNN, Xg-Boost. Select the checkbox for the algorithms you want to use for the captured database. After selecting, the accuracy of each algorithm we selected is predicted. The graph is also arranged as shown in Figure 3. The modal is produced by an algorithm with very high accuracy. Overall, Random Forest has the best accuracy results after pre-processing our database.

2. Step 2: Feature Extraction

After pre-processing, the output is a modal generated algorithm with the highest accuracy. Then from the user interface, input is provided. All 13 parameters namely, gender, age, chol, exang, cp, fbs, trestbps, slope, oldpeak, thal, ca, restecg, thalach. The input provided is transferred to the python backend. By using the modal generated at the pre-processing stage, the input provided is trained. Based on the graph, the patient's stage of illness is known. Therefore, safety precautions should be taken by the patient. It is the best way we can get the result with less risk and less time.

3. Step 3: Classification modeling

In Machine Learning, there are many Classification models which alludes to a predictive modeling by taking the given information as input. Extraction is done based on the models that are having the data classes. A classifier or a classification model predicts categorical classes. Mainly in Machine Learning we perform two types of actions; one is predicting and another one is decision making. In this paper, we have used some of the classification models of Machine Learning. The models we are considered are Gaussian Naive Bayes, Support Vector Machine, Random Forest, K- Nearest Neighbour, Xg- Boost. For each algorithm model, accuracy gets calculated for the given dataset. The dataset gets trained with every algorithm we have check boxed.

a)Gaussian Naïve Bayes: The order of the Naïve Bayes is based on the Bayes Theorem. This approach applies Bayes rules independently for the presence or absence of features. This is a powerful stage for predicting heart disease. It is used to make predictions for each class to separate data sets.

$$P(X_f/C_i) = \frac{P(C_i/X_f)\,P(X_f)}{P(C_i)}$$

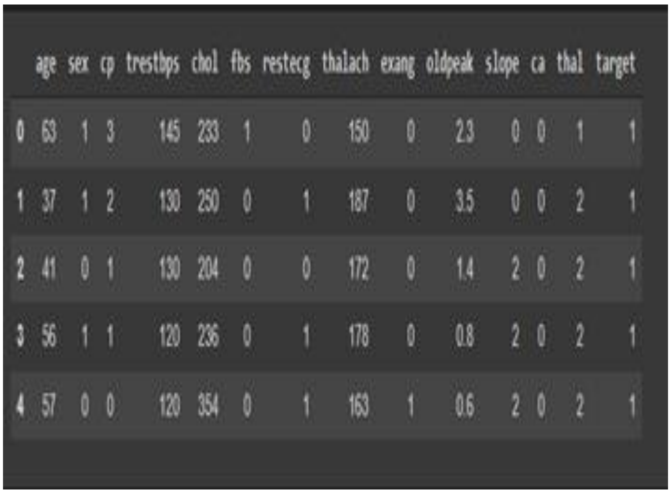
b) Support Vector Machine: Support vector machines a classification method that controls both the line as well non-line data sets. It is a predicted data analysis an algorithm that provides new information parts in one of the marked groups. SVM uses a kernel segmentation tasks for example.

c) Random Forest: Machine Learning algorithms like Random Forest Random Forest can improve the presentation of hazard forecasts by exploiting enormous information repositories to distinguish chance indicators and increasingly complex interactions between them. In Random Forests, we pick an arbitrary determination of highlights for developing the best split. Random Forests Classifier selects a randomly subset of training dataset and then makes a set of decision trees. It decides the votes to decide the final test object class.

d) K nearest neighbor: KNN is a classification algorithm. This is a supervised algorithm. This algorithm is used to extract the knowledge based on the samples distance function and majority of k-nearest neighbors. It checks the whole dataset to find the k nearest instances to the new instance and then output the mode for a classification problem. For some instances KNN algorithm does not go well that is it gets the low accuracy when compared to the others.

e) Xg-Boost: This algorithm uses a gradient boosting framework. For every iteration in this algorithm the error gets reduced as it uses the optimal gradient rather than the methods which uses classical gradients. Xg- Boost models perform the best in laboratory results in each of the cases.

V. Result and Discussion



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Figure 2. DataSet Features Sample

This shows that whether the patient has the heart disease by considering the parameters such as age, gender, chol ,exang, restecg, thal, slope, ca, oldpeak, trestbps, fps, cp and thalach. This experiment is performed by training the dataset containing of 304 records with 13 different parameters. The sample of dataset used is shown in Figure 2.

The distribution of the data plays an important role when the prediction or classification of a problem is to be done. We see that the heart disease occurred 54.46% of the time in the dataset, whilst 45.54% was the no heart disease. So, we need to balance the dataset or otherwise it might get over fit. This will help the model to find a pattern in the dataset that contributes to heart disease and which does not as shown in Figure 3.

After performing all the classification techniques, accuracy of random forest is with 88.52% which is good and higher when compared to other models. The accuracy obtained for all the algorithm models is mentioned below in the table 2. Those are the obtained accuracies after preprocessing our dataset. This is the accuracy chart we have obtained after pre- processing of the classification models we have selected for our Dataset.

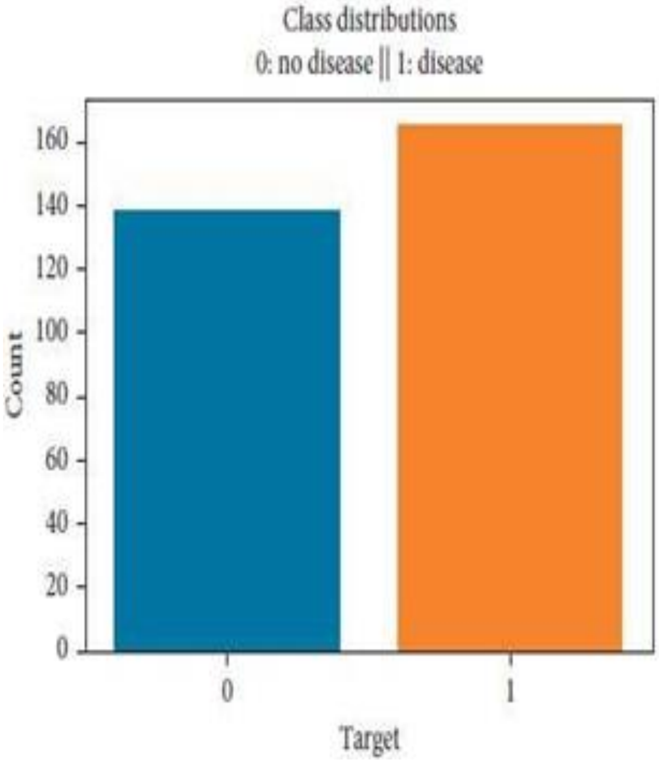


Figure 3.Distribution of Data

Table 2. Machine Learning Algorithms Accuracy

S. No.	Classification Algorithms	Accuracy
1.	Logistic Regression	85.25%
2.	Naïve Bayes	85.25%
3.	Support Vector Machine	81.97%
4.	K-Nearest Neighbors	67.21%
5.	Decision Tree	81.97%
6.	Random Forest	90.16%
7.	XGBoost	85.25%
8.	Neural network	80.33%

In figure 4, we are mentioning the algorithm model and along with the accuracy of each of the algorithm model after training the dataset. The modal is generated with the algorithm having the highest accuracy a sin the process mentioned in figure 4 which is Random Forest. Then in the next phase, the input of 13 different attributes are given from the user interface shown in figure 5.

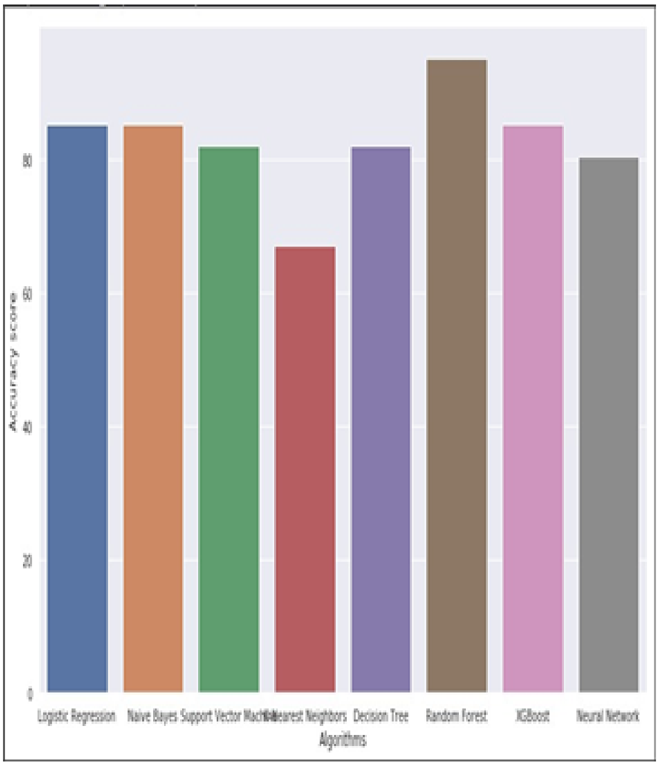


Figure 4. Accuracy score of Different ML Algorithms

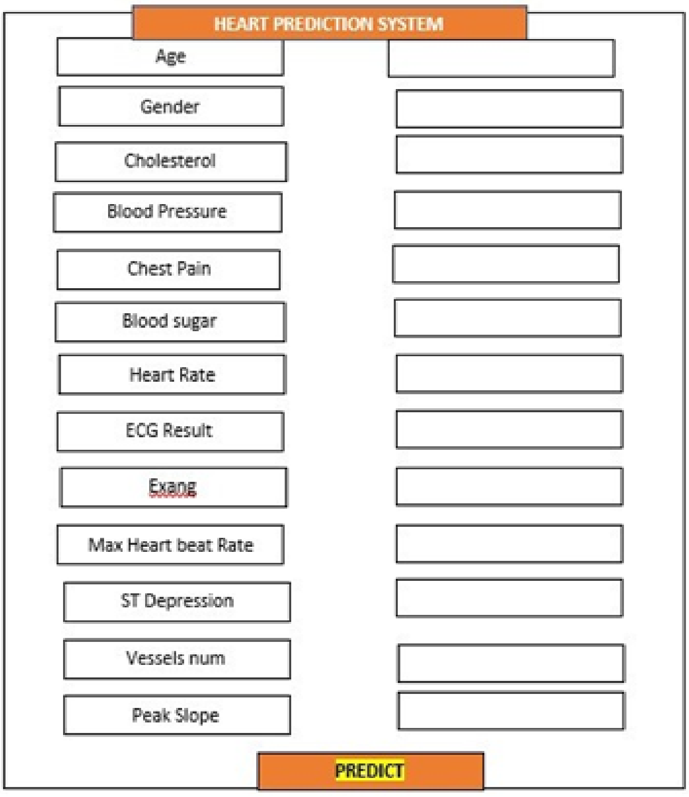


Figure 5. User Interface

VI. Conclusion

In this research paper various machine learning algorithms are used to predict the heart disease. Dataset for heart illness containing its factors have been taken from UCI Machine Learning Repository and further classification models applied on the respective dataset. Our proposed methodology uses various machine learning models namely Support Vector Machine, Random Forest, KNN, Gaussian Naïve Bayes, Xg- Boost algorithms for predicting of heart disease in a short period of time retrieve the results and diminish the expenses for people. We are utilizing these algorithms to improve the standardization.

Table 1. Review of Literature Summary

S.No.	Author	Year	Research Findings
1.	Escamila et al. [2]	2020	Various Medical parameters were used fordetecting theresults with the help of statistical methods suchas DNN and ANN.
2.	Harvard Medical School[3]	2020	Cleveland dataset were usedfor heartdisease prediction withthe help of machine learning classifier and PCA were used for dimensionality reduction.
3.	Zhang et al. [4]	2018	PCA is used with AdaBoostclassifier forthe feature extraction.
4.	Singh et al. [5]	2018	Fisher method and discriminant analysis with binary classifier were used for thefeature extraction.
5.	David et al. [6]	2018	In this paper the performance of various machine learning algorithms areanalyzed usingseveral cross validations.
6.	Chen et al. [7]	2018	Feature reduction wasdone with thehelp of clustering methods.
7.	Kumar [8]	2017	Results ofvarious machine learning
			Algorithms are analyzed to predict heart disease.
8.	Rajagopal et al. [9]	2017	Combination of probabilistic neural network with PCA and kernel PCA were used to perform dimension reduction.
9.	Liu et al. [10]	2017	For feature selection they used Relief F method and heuristic reduced rough set algorithm were used for feature reduction.
10.	Khan et al. [11]	2016	Apply various data mining techniques to analyze unstructured data.
11.	Dun et al. [12]	2016	In this paper hyperparameter techniques are used to increase the accuracy of heart disease prediction.
12.	Imani et al. [13]	2015	In this paper a new weighted training sample method were approach including feature extraction when the data is not enough.
13.	Guidi et al. [14]	2014	SVM and Fuzzy logic were used to predict heart failure with the help of a clinical decision support system.
14.	Santhanam [15]	2013	Regression techniques were used with PCA To extract features.
15.	Ratnasari et al. [16]	2013	Cleveland and UCI dataset are analyzed with the help of feature extraction techniques.
16.	Kamancay et al. [17]	2013	Object recognition was performed with scale-invariant feature transformation.

Table 2.

S. No.	Features	Description	Values
1.	Age	age of a patient in years	Continuous
2.	Gender		1- Male, 0-female
3.	trestbps	Resting blood pressure	continuous
4.	Cp	Type of chest pain	Values are considered between 1 to 4. (1-typical,2-atypical angina, 3- non anginal pain,4-asymptomatic)
5.	fbs	fasting blood sugar	Measured in mg/dland will be Categorized as <=100 normal, between 100 and 125 pre-diabetes, greater than 125 will be categorized as diabetes
6.	restcg	ECG result	0- normal 1- having ST- T2-hypertrophy
7.	Thal	heart rate of patient	It takes following values
			3- normal 6- fixed defect 7- reversible defect
8.	Chol	Serum cholesterol	continuous
9.	Exang	exercise induced angina	Yes, 0- No
10.	Thalach	max heartbeat rate	Continuous values
11.	Oldpeak	ST depression	Continuous
12.	Ca	major vessels number colored by fluoroscopy	Number of major vessels from 0-3
13.	Slope	peak slope	Take value between 1 to 3

References

K. Sharma, "Heart Disease Prediction Using Data Mining Techniques", © IJRAT Special Issue National Conference “NCPC-2016”, pp. 104-106, 19 March 2016.

[2] A. K. G´arate-Escamila, A. Hajjam El Hassani, and E. Andr`es, “Classification models for heart disease prediction using feature selection and PCA,” Informatics in Medicine Unlocked, vol. 19, Article ID 100330, 2020.

[3] Harvard Medical School, “*roughout life, heart attacks are twice as common in men than women,” 2020,

https://www.health.harvard.edu/heart- health/throughout-life-heartattacks-are-twice-as- common-in-men-than-women.

[4] D. Zhang, L. Zou, X. Zhou, and F. He, "Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer," IEEE Access, vol. 6, pp. 28936–28944, 2018.

[5] R. S. Singh, B. S. Saini, and R. K. Sunkaria, "Detection of coronary artery disease by reduced features and extreme learning machine," Medicine and Pharmacy Reports, vol. 91, no. 2, pp. 166–175, 2018.

[6] H. B. F. David and S. A. Belcy , "Heart Disease Prediction Using Data Mining Technques" ICTACT Journal On Soft Computing, 2018, Vol:09, Issue: 01

[7] R. Chen, N. Sun, X. Chen, M. Yang, and Q. Wu, "Supervised feature selection with a stratified feature weighting method," IEEE Access, vol. 6, pp. 15087– 15098, 2018.

[8] S. Kumar, "Predicting and diagnosing of heart disease using machine learning algorithms," International Journal of Engineering and Computer Science, vol. 6, no. 6, pp. 2319–7242, 2017.

[9] R. Rajagopal and V. Ranganathan, "Evaluation of effect of unsupervised dimensionality reduction techniques on automated arrhythmia classification," Biomedical Signal Processing and Control, vol. 34, pp. 1–8, 2017.

[10] X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu,Q. Wang, and Q. Wang, "A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method", Computational and Mathematical Methods in Medicine, Volume 2017, Article ID 8272091, 11pages, DOI: https://doi.org/10.1155/2017/8272091.

[11]S.S.Khan and S.M.K. Quadri, "Prediction of angiographic disease status using rule based data mining techniques," Biological Forum—An International Journal, vol. 8, no. 2,pp. 103–107, 2016.

[12]B.Dun, E.Wang, and S. Majumder, "Heart disease diagnosis on medical data using ensemble learning," 2016.

[13]M. Imani and H. Ghassemian, "Feature extraction using weighted training samples," IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 7, pp. 1387–1391, 2015.

[14] G. Guidi, M. C. Pettenati, P. Melillo, and E. Iadanza, "A machine learning system to improve heart failure patient assistance," IEEE Journal of Biomedical and Health Informatics, vol. 18, no. 6, pp. 1750–1756, 2014.

[15]T.Santhanam and E.P.Ephzibah, "Heart disease classification using PCA and feed forward neural networks," Mining Intelligence and Knowledge Exploration, Springer, Cham, Switzerland, 2013.

[16]N.R.Ratnasari, A. Susanto, I. Soesanti, and Maesadji, "*oracic X-ray features extraction using thresholding-based ROI template and PCA-based features selection for lung TB classification purposes," in Proceedings of the 2013 3rd International Conference on Instrumentation, Communications, Information Technology and Biomedical Engineering (ICICIBME), pp. 65–69, IEEE, Bandung, Indonesia, November 2013.

[17]P. Kamencay, R. Hudec, M. Benco, and M. Zachariasova, "Feature extraction for object recognition using PCA-KNN with application to medical image analysis," in Proceedings of the 2013 36th International Conference on Telecommunications and Signal Processing (TSP), pp. 830–834, IEEE, Rome, Italy, July 2013.

[18]M. Marimuthu, M. Abinaya, K.S. Hariesh, K. Madhankumar, A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach, International Journal of Computer Applications, Vol 181- No. 18, September 2018.

[19]Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techiniques: A Review, ISSN vol 10, No. 7, pp. 2137- 2159.

[20]K.Gomathi Kamaraj, D. Shanmuga Priyaa, Mutli Disease Journal of System and Software Engineering, Dec 2016.

[21]K. Polaraju, D. Durga Prasad, Prediction of Heart Disease using Multiple Linear Regression Model, IJEDR, vol 5, ISSN:2321-9939,2017.

[22]Jaymin Patel, Prof. Teja; Upadhyay, Dr.Samir Patel, Heart Disease Prediction using Machine Learning and Data Mining Technique, IJCSC, vol 7, pp- 129- 137.

[23]Ashwini Shetty A, Chandra Naik, Different Data International Journal of Innovative in Science Engineering and Technology, Vol.5, pp.277-281.

[24]MeghaShahi, R. Kaur Gurm, Heart Disease Prediction Computer Science Technology, vol 6, pp.457-466.

[25]R.Sharmila, S.Chellammal, A conceptual method to enhance the prediction of heart diseases using the data and Engineering, May2018.