INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH AND TECHNOLOGY
ISSN 2582-7358
Impact Factor 6.328
Peer-Reviewed Journal

Effective prediction of Heart disease using Machine Learning Techniques

[1]Rajesh Kumar Pradhan, [2]Priyabrata Sahu, [3]Suvendu Ku Jena

[1]Department of Computer Science Engineering and Applications
Indira Gandhi Institute of Technology, India
[2]Department of Computer Science Engineering and Applications
Indira Gandhi Institute of Technology, India
[3]Department of Computer Science Engineering &Applications
Indira Gandhi Institute of Technology, India

ABSTRACT –

One of the most essential research areas has always been applied in the area of medical development. And now days globally, heart disease (HD), particularly coronary artery disease (CAD) commonly known as atherosclerosis is the leading cause of death. However, the early detection system for cardiac disorders or heart disease, is one of these medical applications to avoid the advanced cases, and reduce treatment costs. As a result, we presented a medical support system for predicting heart disease, which would aid physicians in diagnosis and decision-making. In this work, we used machine learning techniques such as Extra-Tree Classifier, Support Vector Machine, *Logistic Regression,* ***K-Nearest Neighbour and Decision Tree are used to predict heart disease using data from medical files. Using the UCI data set, several experiments have been undertaken to predict HD, and the results show that Extra-Tree Classifier beats both cross-validation and train-test split approaches, with accuracy of 95.00 %, respectively.***

Keywords—Heart Disease, coronary artery disease, atherosclerosis, machine learning, UCI heart disease data set.

I.INTRODUCTION

According to the World Health Organization (WHO), heart disease is one of the leading causes of mortality when the heart is unable to pump oxygenated blood throughout the body. One kind of cardiovascular illness is coronary artery disease (CAD), often known as atherosclerosis (CVD)[1].This condition causes plaque accumulation in the arteries due to cholesterol in the circulation. Because of plaque buildup, this illness results in restricted or obstructed blood vessels and coronary arteries. Cholesterol, calcium, and other chemicals make up this plaque. The plaque restricts blood flow to the coronary arteries as it builds up. As a result, myocardial blood flow declines. This can result in symptoms like angina. Chest, shoulder, abdomen, arms, and neck pain are all possibilities. The amount of oxygenated blood in the body decreases during this pain. Myocardial ischemia is the medical term for this condition. The myocardium tissue dies when the coronary artery is nearly entirely restricted, resulting in a heart attack (myocardial infarction) [2,3].

Developing and implementing a medical diagnostic support system (MDSS) to automate the categorization and prediction of CVD looks to be crucial at this time. In order to give the best clinical suggestions, medical diagnostic research requires a better level of precision and efficiency. Despite the fact that traditional MDSS has proved its capacity to handle the majority of diagnostic problems, it has a lower accuracy factor and is unable to provide a correct diagnosis [4].

Machine learning techniques are currently being used to help produce more accurate HD forecasts based on medical data such as demographic, symptom and assessment, ECG, and laboratory information. Several studies have been undertaken to identify and predict heart disease. Heart disease was studied using machine learning techniques [5].

II. LITERATURE SURVEY

We have included a few selected articles from a literature review on heart disease diagnosis in this section. These studies employed the same well-known databases, which we will compare performance with later.

Machine learning algorithms are used to predict many diseases, and many researchers, such as Kohali et al., have worked on this. [7] Work was done on heart disease prediction using logistic regression, diabetes prediction using support vector machine, and breast cancer prediction using Adaboot classifier, with the results showing that logistic regression has an accuracy of 87.1 %, support vector machine has an accuracy of 85.71 %, and Adaboot classifier has an accuracy of 98.57 %, which is good for prediction.

This neural network integration method is described in and is used to create new models by combining anticipated values from prior models. The accuracy rate was 89.01 percent when compared to the ML algorithm. Another study published in proposed a medical decision support system (MDSS) for predicting cardiac disease using Weighted Fuzzy Rules (WFR). They employed two evaluation scenarios: the first automates the WFR generating process, while the second constructs a fuzzy rule-based MDSS. They used Cleveland's heart disease database to evaluate their MDSS. The best precision value attained by this technology is 62.35 % when compared to the system based on a neural network.

The authors of used Fast Decision Tree (FDT) and C4.5 tree pruning techniques. This method intends to integrate the findings of machine learning analysis into several CAD databases. The classification accuracy was found to be 78.06 %, which is greater than the 75.48 % average classification accuracy of distinct datasets. In 2017, the authors in suggested a Hybrid Neural Network-Genetic (HNNG) to reinforce the neural network's initial weights using a genetic approach. Using the Z-Alizadeh Sani data set and Cleveland's heart disease database, the maximum accuracy percentage is 93.85%.

The author [12] used the Alizadeh Sani data set to create a hybrid method that improved the performance of a neural network using a genetic algorithm and achieved a 93 % accuracy. The author [9] approved a comparative study utilising four different classifiers, including SVM, KNN, and Neural network. Using 14 different attributes but distinct data sets, he reached a high accuracy of 93.02 %. Despite a large body of research, there is no gold-standard model for predicting HD. As a result, there is still scope for improvement. The data set used, the number of attributes and the output class, as well as the algorithm used, all have an impact on the building of the HD prediction model.

Using a Hungarian data set, this research seeks to develop a medical decision support system that can predict the risk level of HD. To find trends in existing HD patient data, a classification approach is suggested. The methodology we utilised is outlined in the next part, along with a brief description of the data set we used. The experiments and various representations of outcomes are presented in Section 3. Finally, in section 4, you'll find the conclusions.

III. PROPOSED METHOD

A. MACHINE LEARNING

Machine Learning is an efficient technology that is based on two terms: testing and training. The system takes training directly from data and experience and then applies this training to different types of needs based on the algorithm required.

There are three type of machine learning algorithms:

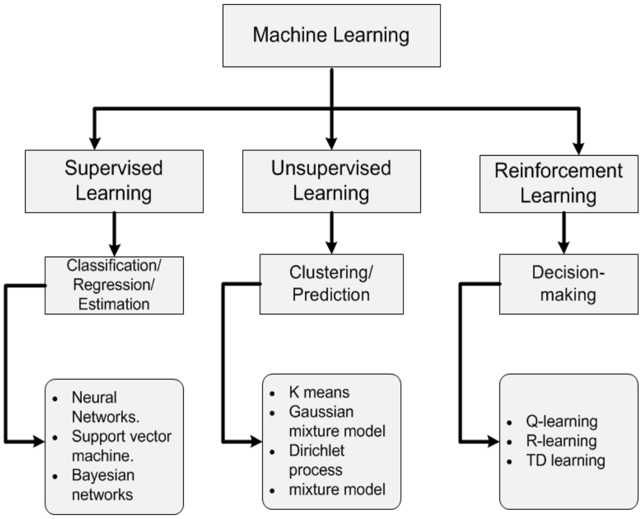


Fig.1 Classification of machine learning

A. Supervised Learning

Supervised learning is described as learning with a proper guide or learning in the presence of a teacher. We get a training dataset that acts as the trainer for prediction on the data set, so there is always a training dataset when testing a dataset. The concept of supervised learning is "train me." The following processes are involved in supervised learning:

- Classification
- Random Forest
- Decision tree
- Regression

The phenomenon of regression is the recognition of patterns and the measurement of the probability of uninterruptible outcomes. The system is capable of identifying numbers, their values, and grouping sense of numbers, which means width and height, among other things. The supervised machine learning algorithms are as follows:

- Linear Regression
- Logistical Regression
- Support Vector Machines (SVM)
- Neural Networks
- Random Forest
- Gradient Boosted Trees
- Decision Trees
- Naive Bayes

B. Unsupervised Learning

Unsupervised learning is described as learning without supervision, in which no trainer is present to provide assistance. When a dataset is supplied to Unsupervised learning, it automatically works on the dataset to uncover patterns and relationships between them, and when fresh data is given, it classifies it and stores it in one of the relationships. Unsupervised learning is founded on the concept of "self-sufficiency."

For example, assume that you have a mixture of fruits such as mango, banana, and apple, and you use Unsupervised learning to classify them into three separate clusters based on their relationship to one another, and whenever new data is received, it is automatically sent to one of the clusters.

Supervisor learning says there are three clusters: mango, banana, and apple, whereas Unsupervised learning says there are three. The procedure for unsupervised algorithms is as follows:

- Dimensionality
- Clustering

There are following unsupervised machine learning algorithms:

- t-SNE
- k-means clustering
- PCA

C. Reinforcement

The agent's ability to interact with the environment and determine the outcome is known as reinforced learning. It is based on the principle of "hit and trial." In reinforced learning, each agent is given positive and negative points, and on the basis of positive points, reinforced learning produces a dataset output. On the basis of positive rewards, reinforced learning trains and test datasets. The procedure for reinforced learning is as follows:

- Q learning
- R learning
- TD learning

B. METHODOLOGY OF SYSTEM

The system's processing begins with data collecting, for which we use the UCI repository dataset, which has been thoroughly confirmed by a number of researchers and the UCI authority.

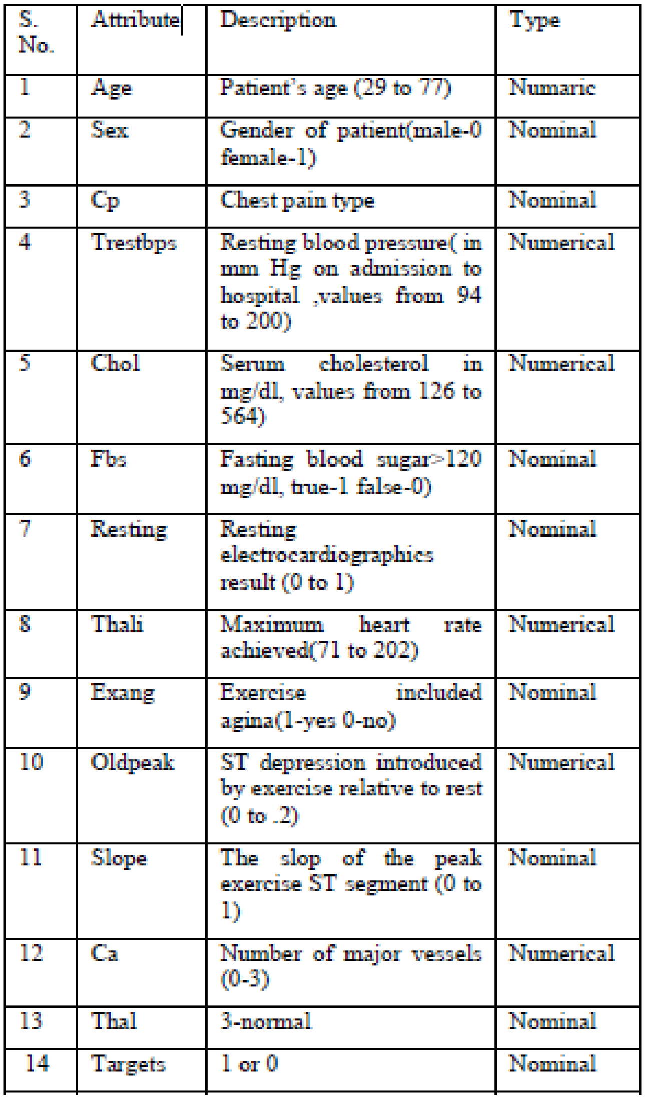
A. Data Collection

The first stage in developing a prediction system is gathering data and settling on a training and testing dataset. In this research, we used 70 % of the training dataset and 30 % of the testing dataset to develop the system.

B. Attribute Selection

Attributes of datasets are properties of datasets that are applied for systems, and for the heart, various attributes such as the person's heart bit rate, gender, age, and many more are shown in TABLE.1 for the prediction system.

TABLE.1 Attributes of the Dataset



S. No.	Attribute	Description	Type
1	Age	Patient's age (29 to 77)	Numaric
2	Sex	Gender of patient(male-0 female-1)	Nominal
3	Cp	Chest pain type	Nominal
4	Trestbps	Resting blood pressure(in mm Hg on admission to hospital ,values from 94 to 200)	Numerical
5	Chol	Serum cholesterol in mg/dl, values from 126 to 564)	Numerical
6	Fbs	Fasting blood sugar>120 mg/dl, true-1 false-0)	Nominal
7	Resting	Resting electrocardiographics result (0 to 1)	Nominal
8	Thali	Maximum heart rate achieved(71 to 202)	Numerical
9	Exang	Exercise included agina(1-yes 0-no)	Nominal
10	Oldpeak	ST depression introduced by exercise relative to rest (0 to .2)	Numerical
11	Slope	The slop of the peak exercise ST segment (0 to 1)	Nominal
12	Ca	Number of major vessels (0-3)	Numerical
13	Thal	3-normal	Nominal
14	Targets	1 or 0	Nominal

C. Data pre-processing

Pre-processing is required for machine learning algorithms to produce prestigious results. For example, the Random Forest technique does not handle datasets with null values, so we must manage null values from the original raw data.
We need to convert some category values to fake values in the form of "0" and "1" for our project.

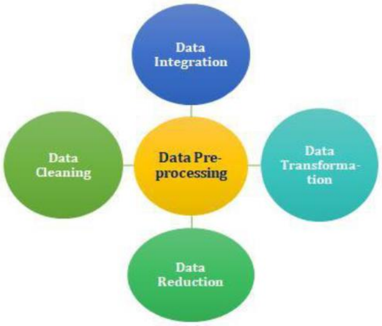


Fig.2 Data pre-processing

D. Data Balancing

Data balancing is necessary for reliable results since we can tell from the data balancing graph that both target classes are equal. The target classes are depicted in Fig.3, with "0" representing patients with heart disease and "1" representing patients without heart disease.

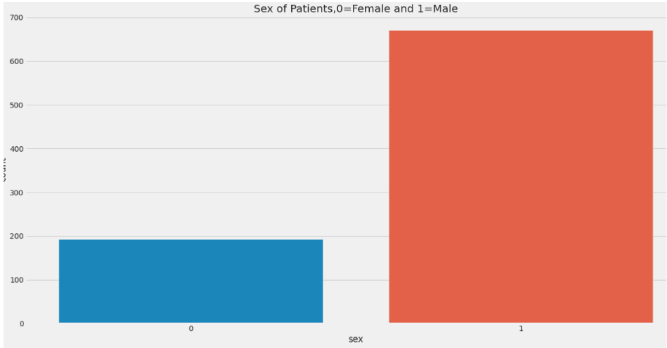


Fig.3 Target class view

E. Disease Prediction

SVM, Naive Bayes, Decision Tree, Random Tree, Logistic Regression, Ada-boost, and Gradient Boosting are machine learning methods used for classification. A comparison of algorithms is conducted for heart disease prediction, and the algorithm with the best results is chosen, and the algorithm with the highest accuracy is employed.

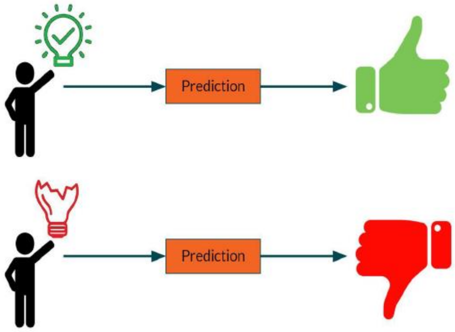


Fig.4 Disease Prediction

IV. MACHINE LEARNING ALGORITHMS

A. Logistic Regression

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

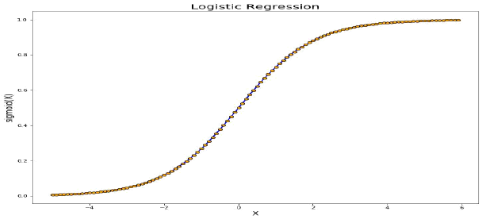


Fig.5 Logistic Regression

B. Decision tree

A decision tree, on the other hand, is a graphical representation of data and a type of supervised machine learning technique.

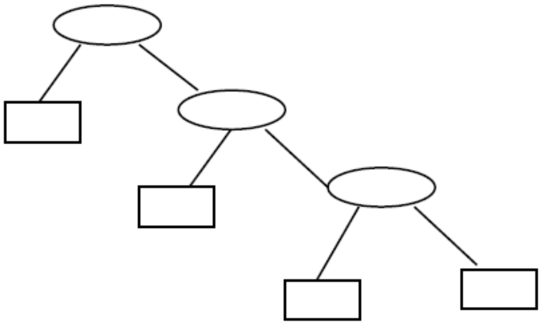


Fig.6 Decision tree

We use the entropy of the data attributes to build the tree, and we draw the root and other nodes based on the attribute root.
Entropy= -Σ Pij log Pij (1)
The entropy of each node is derived using the following equation of entropy (1), where Pij is the probability of the node. The root node is chosen based on the highest entropy calculation, and this procedure is repeated until all of the tree's nodes have been calculated or the tree has been created.
When the number of nodes in a tree is unbalanced, the tree develops an overfitting problem, which is bad for calculations and one of the reasons why decision trees are less accurate than linear regression.

C. Support Vector Machine

It is a type of machine learning technique that works on the concept of the hyper plan, which means that it classifies data by creating a hyperplane between them.
The training sample dataset is (Yi, Xi), where i=1,2,3,.....n and Xi is the ith vector and Yi is the target vector. The number of hyper plans determines the type of support vector; for example, if a line is used as a hyperplane, the method is known as a linear support vector.

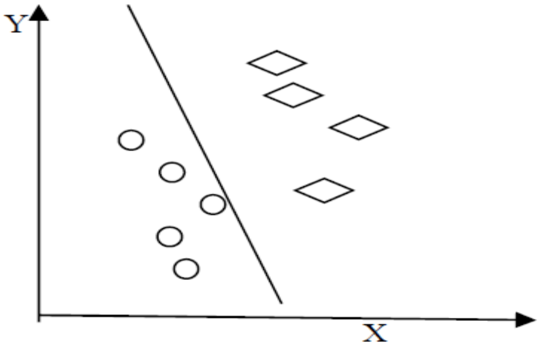


Fig.7 SVM

D. K-Nearest Neighbour

K-Nearest Neighbour Classification is a non-generalizing type of instance-based learning that stores instances of the training data. It finds a predetermined number of training samples that are closest in distance to predict a new location. Weights can also be assigned to new points when they are classified. Each neighbour in scikit-learn can be given a uniform weight or weights proportional to the inverse of the distance from the query point or a user defined weight.

E. Random Forest

A supervised learning algorithm is Random Forest. It is an extension of machine learning classifiers that includes bagging to improve Decision Tree performance. It combines tree predictors, and the trees are dependent on a random vector that is sampled independently. It can be used both for classification and regression.

F. Extra Tree Classifier

Extremely Randomized Trees (or Extra-Trees) is an ensemble learning method. The method creates extra trees in sub-samples of datasets and applies majority voting to improve the predictivity of the classifier. By this approach, the method reduces the variance. The method applies a random threshold for each feature of sub-samples to obtain the best of the thresholds as a splitting rule.

G. Gradient Boosting

Gradient Boosting is a well-known forward learning ensemble method in machine learning. It is an effective method for developing predictive models for regression as well as classification tasks.

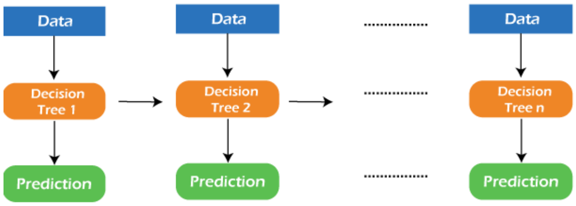


Fig.8 Gradient Boosting

V. Result Analysis

A. About Jupyter Notebook

Jupyter notebook is used as a simulation tool, and it is user-friendly for Python programming projects. Jupyter notebook includes rich text features as well as code, such as figures, equations, links, and many more. Because of the combination of rich text components and code, these documents are ideal for bringing together an analysis description and its findings, as well as executing data analysis in real-time.
Jupyter Notebook is a web-based interactive visual, maps, charts, visualizations, and narrative text application that is free source.

B. Accuracy calculation

The algorithms' accuracy is determined by four values: true positive (TP), false positive (FP), true negative (TN), and false negative (FN) (FN).
Accuracy= (FN+TP) / (TP+FP+TN+FN) (2)
The numerical value of TP, FP, TN, and FN is defined as follows:
TP = Number of individuals with heart disease
TN = Number of individuals with and without heart disease.
FP = Number of individuals who do not have heart disease.
FN denotes the number of individuals who do not have heart disease and those who do have heart disease.

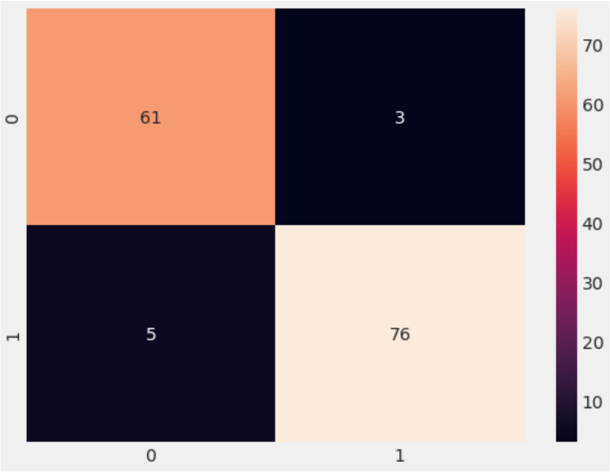


Fig.9 Confusion matrix

C. Result

After using the machine learning technique for testing and training, we discovered that the accuracy of the Extra Tree Classifier is substantially higher than that of other algorithms. Accuracy should be calculated with the help of the confusion matrix of each algorithm, where the number of counts of TP, TN, FP, and FN is given and using the equation (2) of accuracy, value has been calculated and it is concluded that Extra Tree Classifier is the best among them with 95 % accuracy.

VI. CONCLUSION AND FUTURE SCOPE

Because the heart is such an important and critical organ in the human body, and the prediction of heart problems is also a major worry for people, algorithm reliability is one of the parameters used to evaluate the efficiency of the algorithm. The dataset used for training and testing purposes determines the accuracy of machine learning algorithms. We discover that Extra-Tree Classifier is the best algorithm when we compare algorithms using a dataset with the properties indicated in TABLE.1 and a confusion matrix.

In the future, more machine learning approaches will be utilized for the best analysis of cardiac illnesses and for early disease prediction, so that the incidence of mortality cases may be lowered via disease awareness.

References

[1] OMojisola Grace Asogbon; Oluwarotimi Williams Samuel; Shixiong Chen; Pang Feng; Guanglin Li"A Hybrid Approach Based on Non-parametric Attribute Learning Technique and Multi-layer Networks for Congestive Heart Failure Risk Prediction" 2019 IEEE 5th International Conference on Computer and Communications (ICCC),DOI: 10.1109/ICCC47050.2019.9064070

[2] Oumaima Terrada; Bouchaib Cherradi; Soufiane Hamida; Abdelhadi Raihani; Hicham Moujahid; Omar Bouattane"Prediction of Patients with Heart Disease using Artificial Neural Network and Adaptive Boosting techniques" 2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet),DOI: 10.1109/CommNet49926.2020.9199620

[3] Senthilkumar Mohan; Chandrasegar Thirumalai; Gautam Srivastava"Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" DOI: 10.1109/ACCESS.2019.2923707

[4] AniruddhaDutta, TamalBatabyal, MeheliBasu, Scott T.Acton"An efficient convolutional neural network for coronary heart disease prediction" Expert Systems with Applications Volume 159, 30 November 2020, 113408 DOI: 10.1016/j.eswa.2020.113408

[5] E. Nasarian, M. Abdar, M.A. Fahami, R. Alizadehsani, S. Hussain, M.E. Basiri, M. Zomorodi-Moghadam, X. Zhou, P. Pławiak, U.R. Acharya, R.-S. Tan, N. Sarrafzadegan, “Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach,” Pattern Recognition Letters, 133, 33–40, 2020, doi:10.1016/j.patrec.2020.02.010.

[6] M. Akhil, B. L. Deekshatulu, and P. Chandra, “Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm,” Procedia Technol., vol. 10, pp. 85–94, 2013.

[7] 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, 141, 19–26, 2017, doi:10.1016/j.cmpb.2017.01.004.

[8] Pahulpreet Singh Kohli and Shriya Arora, “Application of Machine Learning in Diseases Prediction”, 4th International Conference on Computing Communication And Automation(ICCCA), 2018.

[9] F Ali, S El-Sappagh, SMR Islam, D Kwak, A Ali, Muhammad Imran"A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion" Information Fusion Volume 63, November 2020, Pages 208-222,DOI: 10.1016/j.inffus.2020.06.008

[10] Safial Islam Ayon, Md. Milon Islam, Md. Rahat Hossain"Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques" DOI: 10.1080/03772063.2020.1713916

[11] 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.

[12] Santhana Krishnan J and Geetha S, “Prediction of Heart Disease using Machine Learning Algorithms” ICIICT, 2019.

[13] Aditi Gavhane, Gouthami Kokkula, Isha Panday, Prof. Kailash Devadkar, “Prediction of Heart Disease using Machine Learning”, Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology(ICECA), 2018.

[14] Senthil kumar mohan, chandrasegar thirumalai and Gautam Srivastva, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques” IEEE Access 2019.

[15] Himanshu Sharma and M A Rizvi, “Prediction of Heart Disease using Machine Learning Algorithms: A Survey” International Journal on Recent and Innovation Trends in Computing and Communication Volume: 5 Issue: 8 , IJRITCC August 2017.

[16] 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.