SEYBOLD Report
ISSN: 1533 - 9211

IMPROVED RANDOM FOREST CLASSIFIER FOR PREDICTING HEART DISEASE

Ashish Kumar Gangwar
SCSE, Galgotias University, Greater Noida, India, ashu969027@gmail.com

Priyansh Kamthan
SCSE, Galgotias University, Greater Noida, India
priyansh.kamthan123@gmail.com

Dr. Arvind Dagur
SCSE, Galgotias University, Greater Noida, India
arvind.dagur@galgotiasuniversity.edu.in

Abstract- Heart disease, often known as Coronary artery disease, is the paramount source of death on earth over the past few years. It encompasses a variety of heart-related conditions. It is important to have timely access to reliable, practical, and accurate techniques for disease management and early detection in order to address a number of risk factors for heart disease. In the health maintenance manufacturing, data mining is a frequently utilised approach to handle vast amounts of data. Forecasting heart illness is necessary, researchers analyse vast amounts of complex medical data using a variation several approaches for data excavation and machine learning. The model in this project is based on supervised learning methods including Decision Tree, Decision Tree also Random Forest with K-nearest Neighbor along Support Vector Machine Classifier. It presents numerous heart disease-related features. It utilises the contempory dataset from the Cleveland database from the UCI Coronary artery disease patient history. The gathering has 1026 occurrences with 76 characteristics. Just 14 from the 76 qualities, which were essential to demonstrating how well dissimilar algorithms work, are used in the testing process. Predicting the chance that patients will acquire heart disease is the aim of this research project. According to the results, the Random Forest Classifier Algorithm has the highest accuracy rating.

Keywords— Heart disease Prediction Using Decision Trees, Support Vector Classifiers, Logistic Regression, Random Forest, and KNN similar to excessive cholesterol, obesity, an increase in triglyceride levels, high blood pressure, etc., are to blame for the growth in the risk of heart disease. But, as time passes, There are many hospital patient records and examination data available. The patient's records can be retrieved from a variety of public sources, and further investigation can be done so that a variety of computer technologies can be used to properly analyse the patient's medical data that pinpoint the illness in order to prevent it from becoming fatal.

Artificial _intelligence along machine learning have a remarkable part with health sector. Medical management collects information on numerous health-related issues all across the world. Many machine learning approaches can be used on data to produce insightful perceptions. Yet, the amount of data amassed is enormous, and most of the time, this data may be quite brash. These datasets can be easily navigated using a variety of machine-learning approaches, but they are too overwhelming for human minds to comprehend. As a result, these algorithms have improved and are now quite helpful for specifically predicting the presence or truancy of heart-related disorders. We can identify the disease and distinguish or forcast the consequence using divergent models for deep learning and machine learning. It is possible to alter machine learning models by doing a straightforward genetic data analysis. Models can be instructive for knowledge projection, and medical data can also be transformed and investigated in greater detail for better projection.

I. INTRODUCTION

Heart illness records a spectrum of conditions that dominates your heart. Nowadays, 17.97 million people around globe die every year from Coronary artery disease, which are According to data from the World

II. LITERATURE REVIEW

Systems for predicting heart illness have been created as a consequence of multiple studies conducted in hospitals utilising different machine-learning algorithms.

To increase the accuracy of cardiovascular sickness prediction, Senthil Kumar Mohan et al. [1] suggested a technique called Efficient Coronary artery disease Prediction Using Mixed Breed Learning Approaches.

by using intelligent retrival to identify important elements. A few well-known arranging techniques and a number of highlight combinations are used to generate the expectation model. With an accuracy level of 88.7%, they use a hybrid Random Forest and Linear Model to improve the Coronary artery disease prediction model. (HRFLM). They also received instruction in a variety of data mining methodologies and expectation methods, including K-Nearest Neighbors, Logistic Regression, Support Vector Machine, Neural Network, and Vote Rank. Article titled "Prediction of Heart Disease" by Sonam Nikhar and colleagues [2]. using machine learning methods. This analysis thoroughly examines the decision tree classifier and naive bayes classifier used in this work to predict heart disease. According to certain research that considered the use of predictive data mining approach on a similar dataset, the Decision Tree performs better than the Bayesian classification system.

Heart disease prediction by Dr. Kailas Devadkar, Gouthami Kokkula, Aditi Gavhane, and others utilising machine learning The dataset was used to train and test the proposed system's multi-layer perceptron (MLP) neural network approach. This approach will additionally include one or more hidden layers in between the input and output layers in addition to the input and output layers. Each input layer node is connected to an output layer node via these hidden layers. This connection has weights attached to it. To balance the perceptron, a second similar input with weight b—referred to as bias—is supplied to the node. The connection between the nodes might be either feed-forward or feedback, depending on the circumstance.

III. RELATED WORK

Several studies have been use the UCI Machine Learning dataset to predict cardiac disease. Here are the accuracy levels that have been attained utilising different data mining approaches.

Avinash Golande and colleagues study machine learning (ML) methods that may be used to categorise cardiac disease. The effectiveness of the classification algorithms Decision Tree, KNN, and K-Means was investigated [1]. Decision Trees were shown to have the highest accuracy in the study, and it was decided that by merging other approaches and adjusting its parameters, it would be possible to create a productive algorithm.

A system that combines data mining methods with the Map Reduce approach was suggested by T. Nagamani et al. [2]. The accuracy achieved for the 45 occurrences in the testing set, according to this study, was higher than correctness acquired apply a traditional fuzzy along ANN. Here, the usage of linear scaling and dynamic schema increased the algorithm's accuracy. The ML model developed by Fahd Saleh Alotaibi examines five alternative strategies [3]. Quick Miner, a technique we used, generated outcomes with greater precision than Weka and Matlab's tools. This study assessed the classification accuracy of the methods Decision Tree with Logistic Regression and Random Forest along with Naive Bayes, and SVM. The most accurate algorithm was the decision tree algorithm.

With the purpose of predicting cardiac disease, A survey was conducted by Theresa Princy, R., et al. utilising several classification algorithms. Naive Bayes, KNN (K-Nearest Neighbor), decision trees, and neural networks were the classification methods employed. The accuracy of the classifiers was evaluated for a variety of variables.

IV. PROPOSED MODEL

The proposed study assesses the performance of the four heart disease prediction categorization systems. The investigation's goal is to correctly diagnose cardiac disease in a patient. Information from the patient's medical report is entered by the healthcare provider. The data is included into the prediction model that is used to estimate the likelihood of heart disease. The whole procedure is shown in the figure below.

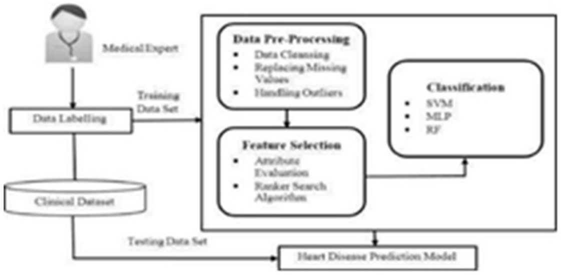
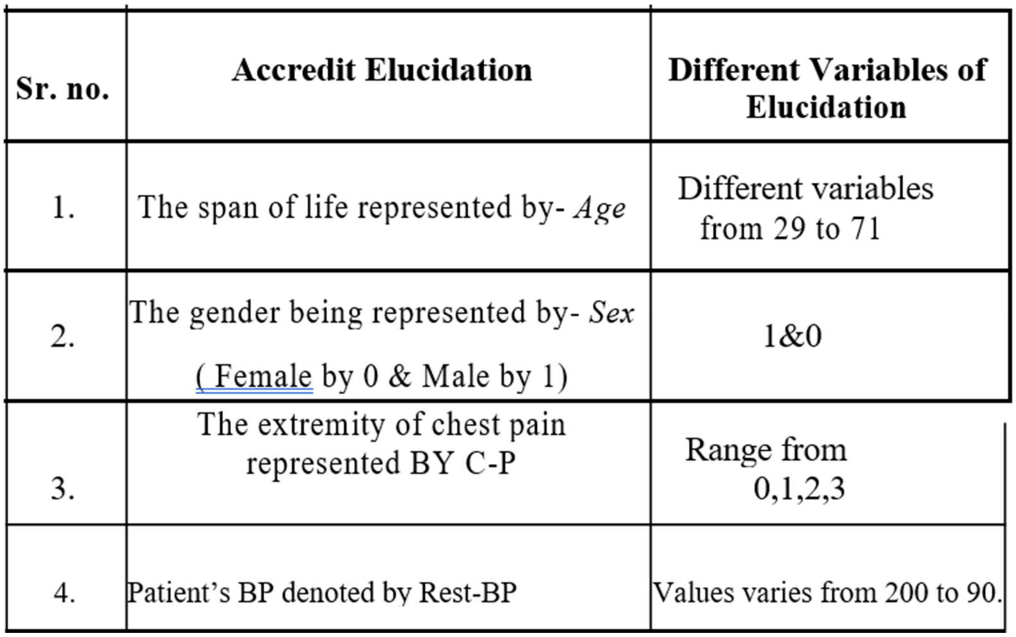


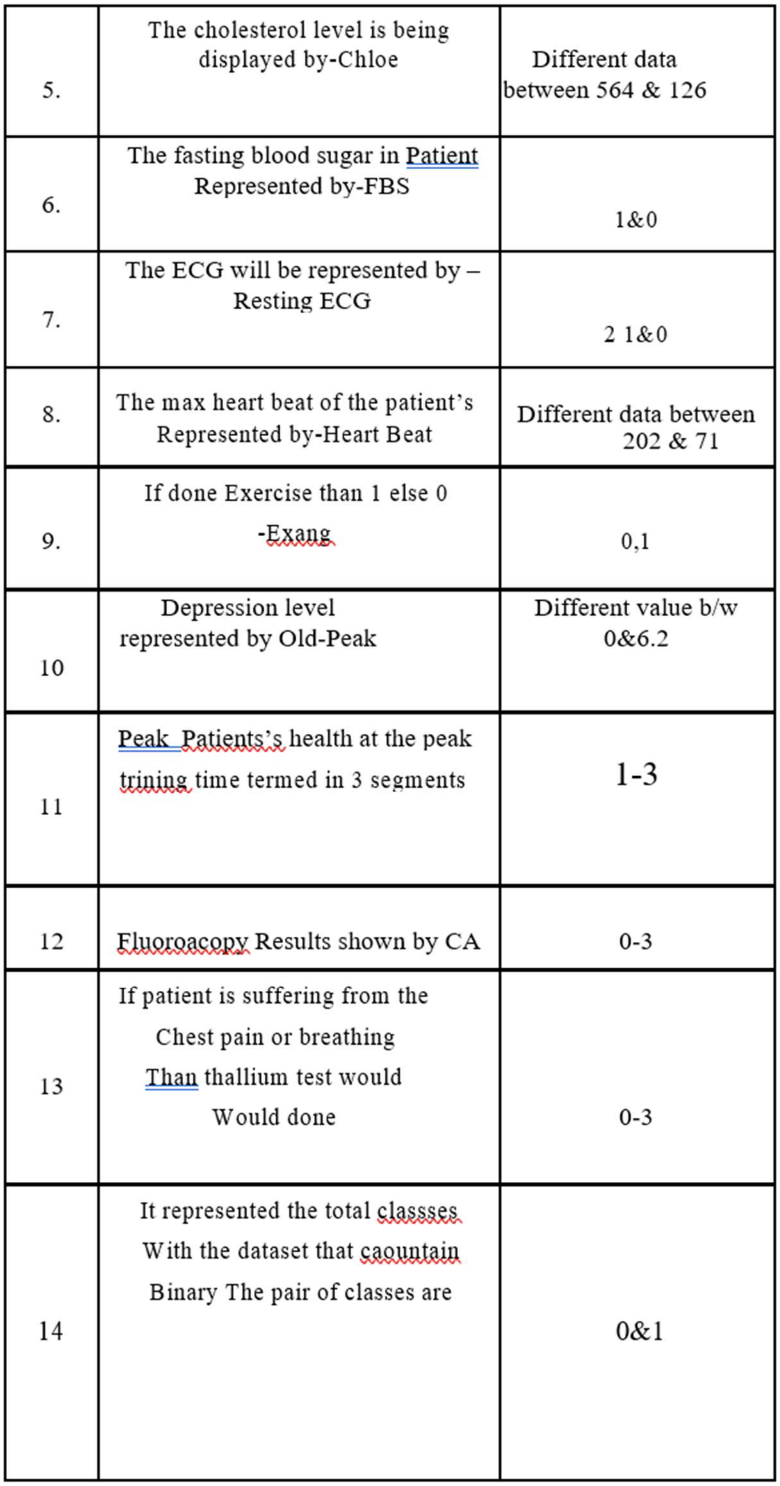
Fig: Architecture Diagram of the proposed model

A. Data Gathering and Preparation - Just the UCI Cleveland dataset from the Coronary artery Disease datafile, a collection of four distinct datasets, was utilised. All published studies only mention using a subset of 14 attributes, despite the fact that this database has a total of 76 properties. As a result, we chose the UCI Cleveland dataset for our investigation, processed and available on the Kaggle website. All 14 criteria used in the suggest research are fully summarised in Table 1 below.:

TABLE I. CHOSEN FEATURES FROM DATASET



Sr. no.	Accredit Elucidation	Different Variables of Elucidation
1.	The span of life represented by- *Age*	Different variables from 29 to 71
2.	The gender being represented by- *Sex* (Female by 0 & Male by 1)	1&0
3.	The extremity of chest pain represented BY C-P	Range from 0,1,2,3
4.	Patient's BP denoted by Rest-BP	Values varies from 200 to 90.



5.	The cholesterol level is being displayed by-Chloe	Different data between 564 & 126
6.	The fasting blood sugar in Patient Represented by-FBS	1&0
7.	The ECG will be represented by – Resting ECG	2 1&0
8.	The max heart beat of the patient's Represented by-Heart Beat	Different data between 202 & 71
9.	If done Exercise than 1 else 0 -Exang	0,1
10	Depression level represented by Old-Peak	Different value b/w 0&6.2
11	Peak Patients's health at the peak trining time termed in 3 segments	1-3
12	Fluoroacopy Results shown by CA	0-3
13	If patient is suffering from the Chest pain or breathing Than thallium test would Would done	0-3
14	It represented the total classses With the dataset that caountain Binary The pair of classes are	0&1

B. Classification- For the various ML algorithms, such as the classification techniques employed by Random Forest with Decision Tree along Logistic-Regression(LR), Naïve-Bayes(NB), the attributes listed in Table 1 serve as the input. The given dataset is classified into a given dataset made up of 80% of it and a test dataset made up of 20% of it. A model is trained using a dataset known as the training dataset. The trained model's fulfillment is calculated using a testing dataset. The effectiveness of each method is assessed and calculated in accordance with a variety of factors, such as correctness, exactness, recall, and F-calculate result. The many algorithms examined in the study given below.

A. Random Forest- Using Classification and regression using Random-Forest algorithms(RFA) are both done. It creates a records tree and uses that to inform its predictions. The Random Forest method may provide the same outcomes when applied to large datasets, even if a substantial fraction of the data records are missing. The tree's generated prototypes can be saved and used with further data. Create a random forest in the first step of Random-forest, and after that use the classifier produced in the primary stage to generate a projection.

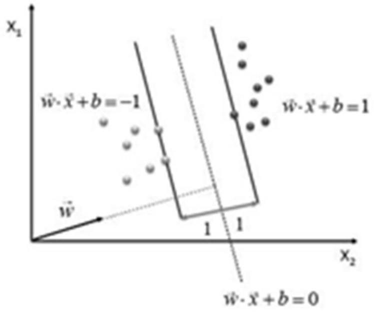
$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \quad Entropy = \sum_{i=1}^{C} - p_i * \log_2(p_i)$$

pi : Relative Frequency

B. Decision Tree-The outside branches of the Decision Tree algorithm indicate the dataset attributes, while the centre node represents the outcome. Because are they rapid, trustworthy, calm to comprehend, and need very small data composition, decision trees are often utilised. The class label prediction is determined by the decision tree's root. The root attribute's value is compared with that of the documentation attribute. Depending over the outcome of the differentiation, the Corresponding branch is followed to the appropriate value before jumping to the next node.

C.Logistic Regression- The binary classification problems are where this classification strategy is most commonly used. In logistic regression, the planning purpose is used to compress the product of a straight equation across 0 -1 rather than fixing a forward path or hyperspace. LR is useful for classifying data since it has 13 independent parameters.

D.Support Vector Machine- Issues with categorization and regression are both handled. The SVM method seeks to define the best line or decision boundary that can divide n-dimensional space into classes in order to quickly categorise fresh data points in the future. This best option boundary is known as a hyperplane. SVM is used to choose the extreme vectors and points that contribute to the hyperplane. The SVM method is built on support vectors, which are utilised to represent these extreme circumstances.


x_1
$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=1$
$\vec{w}$
1
1
x_2
$\vec{w}\cdot\vec{x}+b=0$

$$\boxed{\max \frac{2}{\|w\|}}$$

s.t.

$$(w \cdot x + b) \geq 1, \forall x \text{ of class } 1$$
$$(w \cdot x + b) \leq -1, \forall x \text{ of class } 2$$

E. K- Nearest Neighbors- One of the most basic machine learning algorithms is KNN, which uses the supervised learning method. The K-NN method makes the assumption that the new case and the existing cases are comparable, and it put the new instance with the classification that will most like existing classification. A new facts details are clandestine using the K-NN after all the current data has been saved, an algorithm based on similarity will be used. This suggests that by using the K-NN approach, new data may be categorised rapidly and reliably..

F. IMPROVED RFC- - Using Random Forest Classifier may be used to resolve classification or regression issues. The bootstrap data sample is selected from a training set with replacement, which is the basis for every decision tree in the e group that makes up the random forest method sample. Random forest classifier algorithm can be improved by increasing the number of decision tree. In the random forest classifier, we made an effort to add more trees and get rid of any attributes that weren't crucial to our model. Also we changed the train test split to the ratio of 4:1. The model's accuracy has increased to 95.36% as a result of all these efforts.

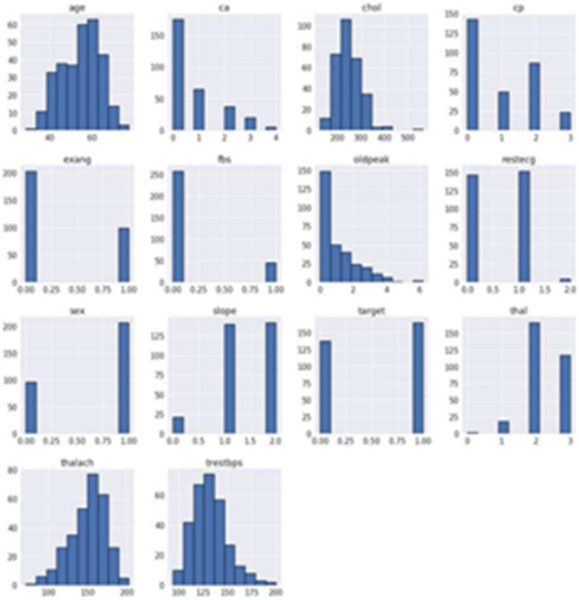
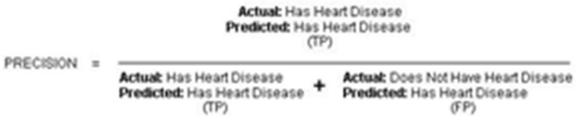


FIG. GRAPH OF PARAMETERS

V. RESULTANDANALYSIS

The results of using Random Forest, Decision Tree, Logistic Regression, Support Vector Machine, and K-Nearest Neighbor are displayed in this section. Using Correctness score, Exactness (P), Reallocate (R), and F- quantify, the algorithm's efficiency is evaluated. The correct measure of positive analysis is provided by the precision metric. The quantity of actual correct positives is defined by recall. The F-measure evaluates precision.



$$\text{PRECISION} = \frac{\begin{array}{c}\textbf{Actual:}\ \text{Has Heart Disease}\\ \textbf{Predicted:}\ \text{Has Heart Disease}\\ \text{(TP)}\end{array}}{\begin{array}{c}\textbf{Actual:}\ \text{Has Heart Disease}\\ \textbf{Predicted:}\ \text{Has Heart Disease}\\ \text{(TP)}\end{array} + \begin{array}{c}\textbf{Actual:}\ \text{Does Not Have Heart Disease}\\ \textbf{Predicted:}\ \text{Has Heart Disease}\\ \text{(FP)}\end{array}}$$

$$\text{RECALL} = \frac{\text{TRUE POSITIVES (TP)}}{\text{TRUE POSITIVES (TP)} + \text{FALSE NEGATIVES (FN)}}$$

•TP True Approving: the test is positive and the suffer has the condition.
•Incorrect positives (FP) are when a test results are positive despite the patient not having the condition.
•TN True rejection: the test is bad and the patient does not have the condition.
•FN False Negative (test result is negative despite the patient having the illness).

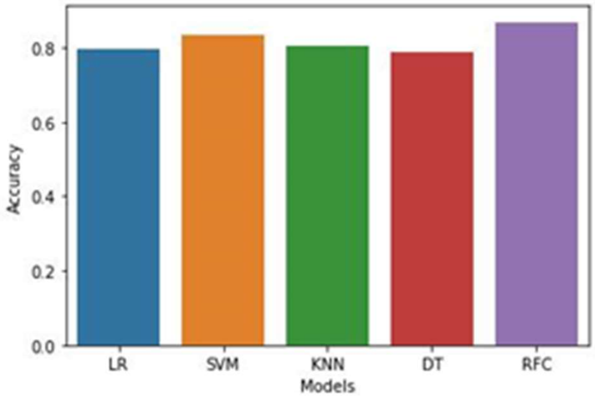
S. no.	Algorithm	Precision	Recall	F-measure
01.	Random Forest	0.87	0.85	0.86
02.	Decision Tree	0.83	0.82	0.82
03.	Support Vector Machine	0.84	0.83	0.83
04.	Logistic Regression	0.79	0.81	0.80
05.	K- nearest Neighbor	0.80	0.81	0.82

Use is made of the pre-processed dataset. in the exercise to run tests, and with the aforementioned techniques are investigated and used. The success indicators first presented are determined by the uncertainty matrix. The uncertainty matrix explains how the algorithm behaves. The uncertainty matrix that proposed model along various approaches are produced. The precision ratings for the methods of Logistic-Regression(LR) or Decision-Tree(DT), Random-Forest(RF), KNN, and Support-Vector-Machine(SVM) are given in the following table:

TABLE II. ACCURACY OF THE ALGORITHMS USED

SI.NO	Model	Training Accuracy %	Testing Accuracy %
1	LR	88.79	86.81
2	K-neares- neighbor	86.79	86.81
3	SVM	93.40	87.91
4	Classiier	100.00	78.02
5	Random Forest Classifier	100.00	82.42
6	Improved RFC	100	95.36

FIG. COMPARISON OF ALGORITHMS



IV. Discussion

Various total systems based on machine learning were discovered through an early study of the literature. The first writers were able to split up the primary difficulty into four components: smart contracts, access judriction, scalability, and patients' live fitness monitoring, despite the fact that the outcomes of the apps greatly diverged from the proposed answer. The effective patient information management systems can be used as a comparison to see how the suggested characteristics compare. The patient's live health tracking used to show by using a wearable tool, despite the failure to transmit the device. There is an alternative tactic that can be used to achieve the goal without endangering the requirements of the user. A class model that is purely based on sensitivity levels was developed using the information gathered from the poll, conversations about smartphones, and staff interviews.

Consequently, secret facts need to be wrapped up for storing, transmission, and change.

IV. CONCLUSION

This research offers a thorough understanding of machine learning methods for categorizing cardiac disorders. In order to forecast the treatment that can be given to patients, classifiers play a key role in the healthcare business. In order to identify the effective and precise methods, the existing methodologies are examined and contrasted.

Machine learning approaches dramatically increase the accuracy of cardiovascular risk prediction, allowing for the early diagnosis of patients who can then receive preventative care.

Conclusion Machine learning algorithms offer a great deal of guarantee for predicting cardiovascular or heart problems. Each of the above mentioned algorithms has performed fantastically in certain circumstances and appallingly in others.

VII.REFERENCES

[1] Avinash Golande, Pavan Kumar T, Heart Disease Prediction Using Effective Machine Learning Techniques, International Journal of Recent Technology and Engineering, Vol 8,pp.944-950,2019.

[2] T.Nagamani, S.Logeswari, B.Gomathy, Heart Disease Prediction using Data Mining with Mapreduce Algorithm‖, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.

[3] Fahd Saleh Alotaibi,‖ Implementation of Machine Learning Model to Predict Heart Failure Disease‖, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.

[4] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, Design And Implementation Heart Disease Prediction Using Naives Bayesian, International
Conference on Trends in Electronics and Information(ICOEI 2019).

[5] Theresa Princy R,J. Thomas,'Human heart Disease Prediction System using Data Mining Techniques',International Conference on Circuit Power and Computing Technologies,Bangalore,2016.

[6] Nagaraj M Lutimath,Chethan C,Basavaraj SPol.,'Prediction Of Heart Disease using Machine Learning', International journal Of Recent Technology and Engineering,8,(2S10), pp 474-477, 2019.

[7] A. K and A. S. Singh, "Detection of Paddy Crops Diseases and Early Diagnosis Using Faster Regional Convolutional Neural Networks," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2021, pp. 898-902, doi: 10.1109/ICACITE51222.2021.9404759.

[8] UCI, ―Heart Disease Data Set.[Online].Available (Accessed on May 1 2020): https://www.kaggle.com/ronitf/heart-disease-uci.

[9] Sayali Ambekar, Rashmi Phalnikar,―DiseaseRisk Prediction by Using
Convolutional Neural Network‖,2018 Fourth International Conference on Computing Communication Control and Automation.

[10] C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, ―Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field,‖ in Machine
Learning Paradigms, 2019, pp. 71–99.

[11] Jafar Alzubi, Anand Nayyar, Akshi Kumar. "Machine Learning from Theory to Algorithms: An Overview", Journal of Physics:
Conference Series, 2018

[12] Fajr Ibrahem Alarsan., and Mamoon
Younes
_Analysis and classification of heart diseases using heartbeat features and machine learning algorithms',Journal Of Big Data,2019;6:81.

[13] Internet source [Online].Available (Accessed on May 1 2020): http://acadpubl.eu/ap

[14] A. Chanchal, A. S. Singh and K. Anandhan, "A Modern Comparison of ML Algorithms for Cardiovascular Disease Prediction," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2021, pp. 1- 5, doi: 10.1109/ICRITO51393.2021.9596228.

[15] A. K, D. D, A. Lakhanpal, K. Manoj Sagar, K. Murugan and A. Shanker Singh, "Discover Pretend Disease News Misleading Data in Social Media Networks Using Machine

Learning Techniques," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2021, pp. 784-788, doi: 10.1109/ICACITE51222.2021.9404648.