Research Article

An Exploiting Machine Learning Technique for Predicting Disease

S.Usha[1], Dr.S.Kanchana[2]

Research Scholar, Department of Computer Science, SRM Institute of Science & Technology, Kattankulathur
Assistant Professor, Department of Computer Science, SRM Institute of Science & Technology, Kattankulathur

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 23 May 2021

Abstract: -
In the field of cardiology, coronary illness is assuming a crucial part since it is a significant reason for death everywhere on the world. In prior days ECG and PCG are utilized to avert and forecast coronary illness. It is exceptionally hard for some individuals to recognize the sickness on the grounds that the expense of the therapy is high. The doctor's job is to prognostic the coronary illness with the assistance of accessible information. The proposed works predict coronary illness by using Machine Learning Techniques. This technique helps doctors to prognostic coronary illness in the most straightforward way. In this work, the following classification algorithm Decision Tree, KNN, Logistic Regression, Naïve Bayes, Support Vector Machine of machine learning are used to predict heart disease. This methodology assists doctors with controlling the model which can improve the precision of prognostic. This model enhances the clinical data and oddity the best calculation by investigating its exactness result and furthermore creates generous mindfulness in the forecast of coronary illness.

Keywords: Decision Tree, KNN, Logistic Regression, Naïve Bayes , Predication, Support Vector Machine.

INTRODUCTION

The human body is comprised of different organs, all of which have their capacities. The heart is one such organ that siphons blood all through the body and if it doesn't do as such, the human body can have lethal conditions. One of the primary reasons for mortality today is having a heart disease. Heart disease can influence the functioning cycle of the heart [1]. According to WHO, heart disease will lead people to more death. The reason for the increase is there will be a lack of medical resources or not having proper medical care. It is very important to monitor the disease constantly and detect heart disease to avoid high risk. Heart illnesses are relied upon to be the primary justification 35 to 60 percent of complete passing anticipated worldwide by 2025[2]. It is very important to prevent heart diseases by diagnosing or predict heart diseases. It includes *a* variety of *tests* to perform the diagnosis. The test includes auscultation, ECG, blood pressure, cholesterol, and blood sugar. These tests are frequently performed when a patient's condition may be critical and he or she must start taking the medication immediately, so it becomes important to prioritize the tests [3]. It is a part of researchers to provide a good predicting model by use of the best data-driven system. This helps people to live healthy lives.

At present the machine learning is said to be an arising field because of the increasing medical data. It very helpful technique for the researcher to secure data from a gigantic extent data, which is extremely hefty for man and now and again inconceivable[1]. As ML algorithm penetrates clinical cardiology, they might be conveyed in numerous regions. This may help the front office plan various patients with the fitting measure of time-dependent on their Electronic Health Record (EHR) information, or help essential consideration doctors better guide references to the cardiologist's office. ML will help in distant testing and checking of patients, directing the information securing that will be shipped off a centre or medical clinic.

I. Related Work

In this exploration, machine-learned prescient models were formed which be determined by the procedures of the hypothesis of nonlinear elements for extricating significant highlights from the OVG and PPG signal information[2].

A clinical decision support system (CDSS) suggests a chance to lessen clinical blunders just as to improve patient security. Perhaps the main uses of such frameworks is in analysis and treatment of heart

disease (HD) on the grounds that measurements have shown that heart disease is one of the main sources of passing everywhere on the world. Data mining techniques have been extremely powerful in planning clinical support systems on account of its capacity find covered up examples and connections in clinical information. With the help of various data mining techniques the CDSS predicted the heart disease. [3].

Heart patients are becoming expediently attributable to inadequate wellbeing mindfulness and terrible utilization ways of life. Subsequently, it is fundamental to have a structure that can strongly perceive the commonness of heart dis-ease in large number of tests momentarily. Different machine learning model was assessed for expectation of heart disease. It achieved 85% of highest accuracy by using Logistic regression with 89% of sensitivity and 81% of specificity [4].

Health care system is very rich in getting information. The wealth of data available in this system provides better accuracy to predict the disease. Number of methods has been directed to analyze the presentation of prescient data mining strategy on the equivalent dataset. The output reveals that decision tree and Bayesian classification produce same accuracy when compare with other models. Genetic algorithm decreases the size of the data for getting the optimal subset which improve the accuracy of decision tree and Bayesian Classification [8]. The classifier algorithm Naïve Bayes and Decision tree are used in the system to predict heart disease[6].

A prediction system is very essential for giving alertness about the diseases. Accuracy was calculated by using machine learning algorithms. The algorithm used in this system is KNN, Decision Tree, linear regression, SVM [7]. A framework that capably stumbling to predict the risk level of patients. It helps no specialized doctors by generating various rules. The generated rules generated are Original Rules, Pruned Rules, and Rules without duplicates, Classified Rules, Sorted Rules, and Polish [8].

N2Genetic optimizer is a new optimization technique that provided 93.08% accuracy and 91.51% of F1 score. This technique enriched the performance of the machine learning algorithm by testing the three types of SVM and normalized the data preprocessing. To achieve the target a genetic algorithm and particle swarm optimization, combined with stratified 10-fold cross-validation, were utilized twice [12].

Accuracy results have been improved by using the attributes age, blood pressure, thickness of the artery, etc. in the algorithm SVM and PCA. The proposed system is used to diagnose heart attack risk. From the experimental results, it is concluded that SVM provides the highest accuracy [9].

II. Methodology

A. Data Collection and Description

The proposed work architecture shows the process of the data that escort to find the accuracy and select the best fit model. The work started with data collection. The data set has been taken from the Kaggle competition platform. Dataset encompasses 70000 patient records of subjects cardiovascular diseases are present or not. Each data is articulated with 11 features which can be described as subjective, objective, and examination features. The objective features afford accurate information that entails age, height, weight, and gender. The examination features are consequences of clinical assessments containing the systolic and diastolic blood pressures just as the groupings of cholesterol and glucose. The subjective features are data given by the subjects including the situation with smoking, liquor taken, and active work.

B. Data Pre-processing

Initially, Electronic Health Data is given as input in the proposed work. In this system, electronic health data is given as input which is an initial step to process the proposed work. Data pre-processing is the most needed to make sure the dataset in the machine learning model. Various steps are included to perform pre-processing. The data set information analyses the data required to predicting the disease.

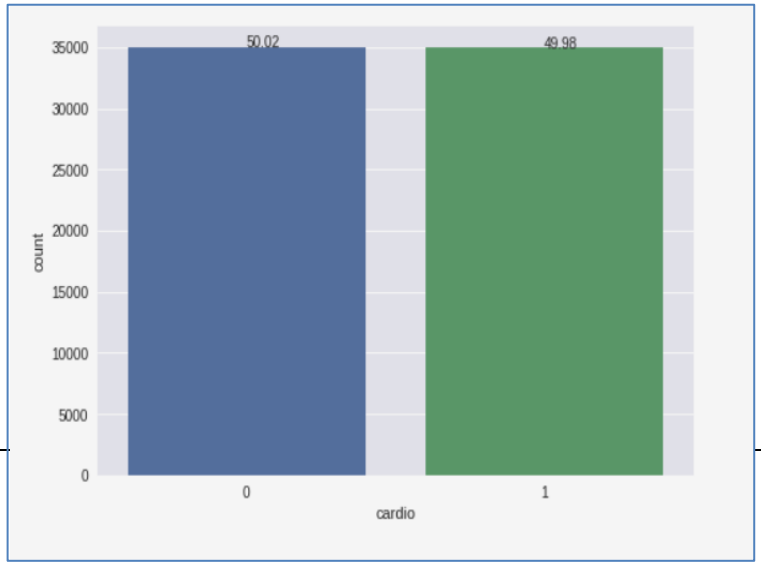

Figure 1 Data Balance Graph

Statistical Details comes up with statistical information is always in the numerical format. The values represent in the mean would use the entire attribute in the given dataset. It is very essential to balance the data. Fig. 1 is a data balance graph which says the target classes where to contain "0" if the patients have heart diseases patient and "1" if no heart diseases patients. The series of dataset attributes and codes are shown as histogram in Fig. 2

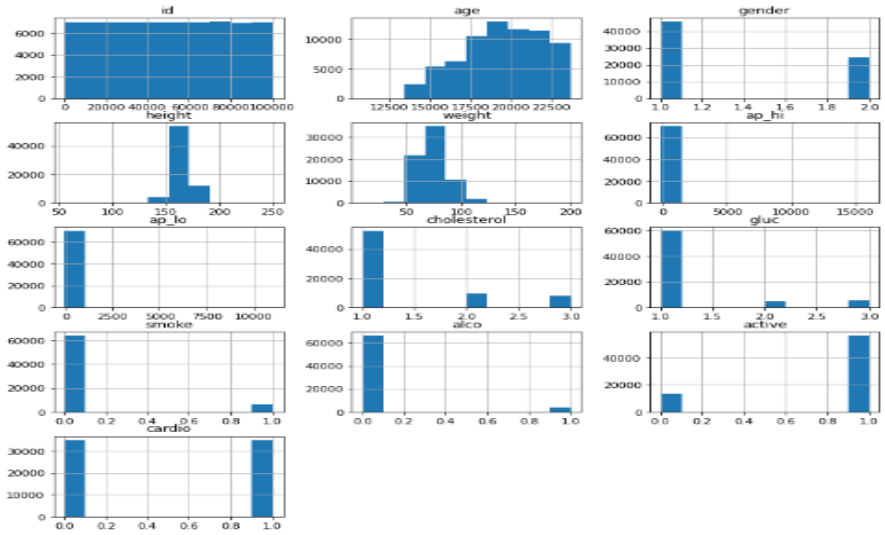

Figure 2 Visualization of Histogram

The dataset has been shaped by removing the duplicate records. The attribute age is given in days. The value of age attribute is change into years for better arrangement because it was in days. The changed data check the connection with the objective variable. Fig. 3 shows that the people over the age of 54 are bound to have infected then beneath, additionally Men underneath 50 are bound to have been determined to have coronary illness than females which affirms our supposition, despite the fact that the thing that matters isn't excessively exceptional.

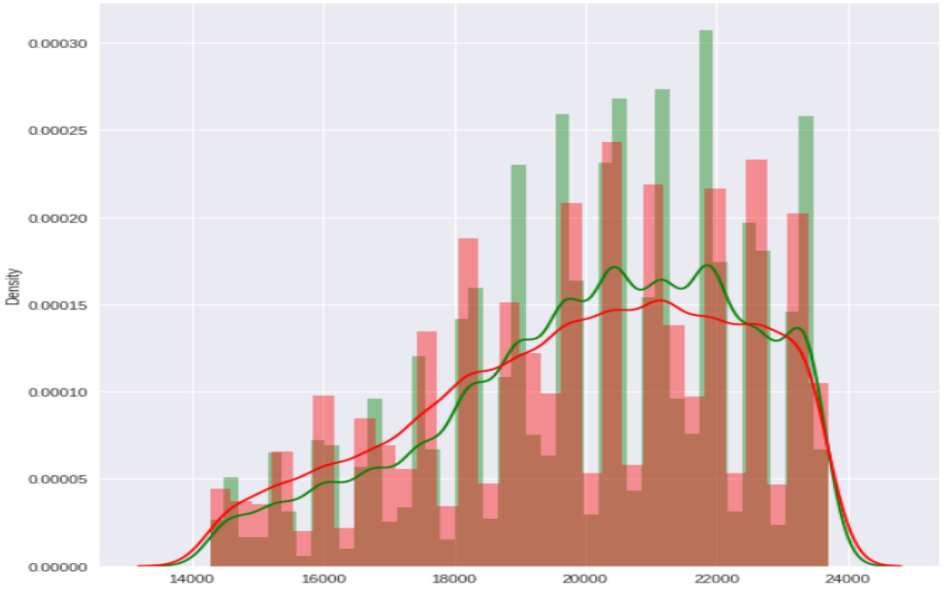

Figure 3 Disease count distribution by gender below the age 54

In the fig. 4 the graph shows that each individual person with less than 4 foot in tallness is generally matured over 40 and has a load above 40kg for the most part. This certainly affirms that they are not youngsters. Presently for the proposed work removed such records for the dataset.

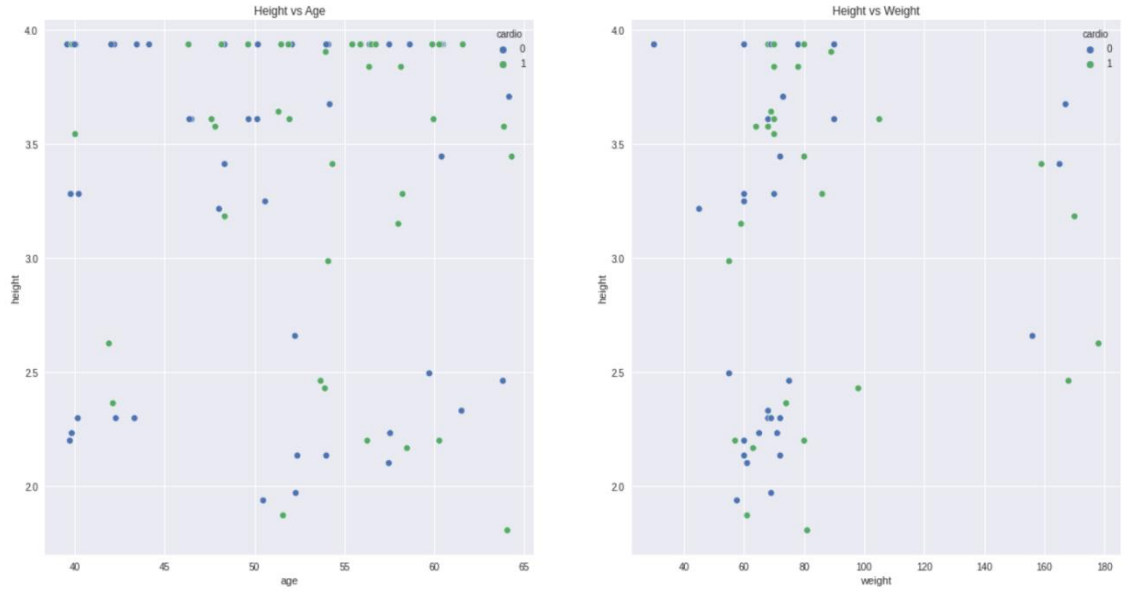

Figure 4 Visualization of dataset in Height and Weight basis

A heat map represents the correlation among the data. The data which establish out is plotted here in Fig 5.

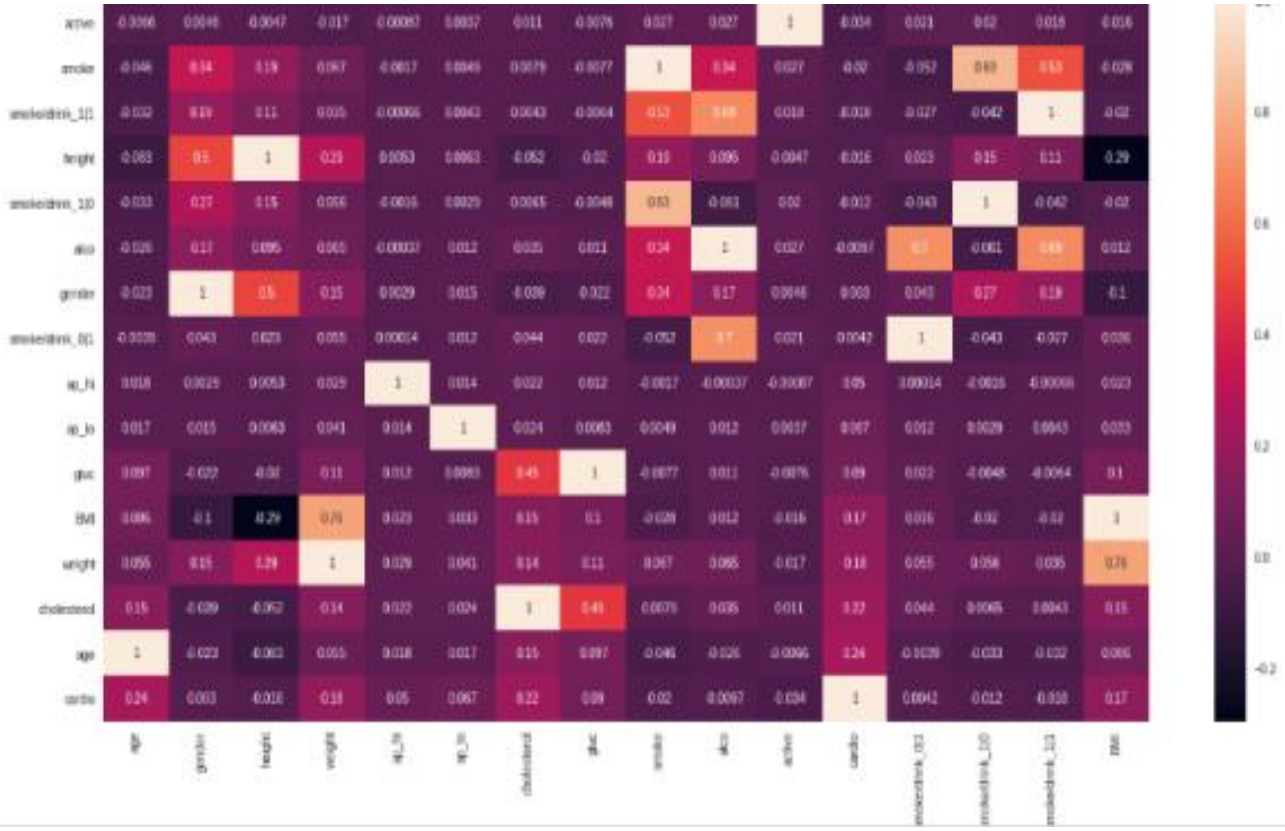

Figure 5. Heat Map Correlation

The positive correlation with the target attribute shows the heat map clearly. Checking the correlation is done here. The data set has been apportioned into two parts; one is training data which contain 75% of the whole data set and another is testing data which remaining dataset. Once data is prepared, the algorithm techniques are implemented and then the confusion matrix has been created. The outcomes have been formed in terms of the accuracy of the algorithm. The confusion matrix is used to find the algorithm accuracy. It is calculated by using classification accuracy. This matrix tells exactly how the instance of each class is assigned. Confusion Matrix Layout is shown in the Fig6.

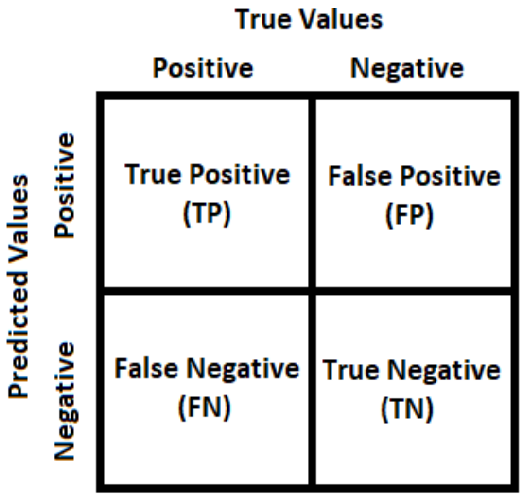

Figure 6 Confusion Matrix Layout

The formula to calculate the accuracy of the algorithm is:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Figure 7 Accuracy Formula

The results obtained after applying the algorithms is as follows
TP (True Positive): Number of records classified as true while they were actually true.
FN (False Negative): Number of records classified as false while they were actually true.
FP (False Positive): Number of records classified as true while they were actually false.
TN (True Negative): Number of records classified as false while they were actually false

C. Algorithms Used

***1.*Decision Tree**

Regression and classification problems are solved by the Decision Tree algorithm. It is one of a supervised learning algorithm. The main aim of this algorithm is to form a training model. The training model can predict the target variable by applying learning simple decision rules. Since the decision classifier is said to be a tree-structured classifier, where the features of a dataset is represented as an internal node, decision rules represent as a branch and each leaf node represent the outcome.

.

***2.* Support Vector Machine**

A Support vector machine is a supervised technique that analyses the data. In this technique, the patterns are discovered. This algorithm is used when the class is two. The data in the class find the best hyper plane which separates all points of one class from another. Support vector machine is very useful to find the solution for complex problem.

***3.* Naïve Baye's**

Naïve Bayes is a powerful supervised algorithm mainly focused to predict the modeling. The name Naïve is to suppose the incidence of a certain feature is independent of the incidence of other features and Bayes is supposed to represent Baye's theorem. Baye's theorem gives the conditional probability of an event A given another event B has occurred.

$P(A|B) = (P(B|A) * P(A)) / P(B)$

Where

P(A|B) : Conditional Probability of A given B.
This is also said as posterior probability

P(A) : Probability of event which is seen after evidence. This is said as prior probability.
P(B) : probability of the data
P(B|A) : Conditional probability of B given

It is not difficult to work with a Naive Bayesian model. Iterative boundary assessment in this model makes to diagnosis heart patients in the field of clinical science. In this classifier a small amount of training data imprecise the parameters that are necessary for classification. Since sovereign factors are accepted, just the differences of the factors for each class should be resolved and not the whole. It tends to be utilized for both parallel and multi class grouping issues [10]

***4.*Logistic regression**

Logistic regression is exploiting for classification problem. It is one of machine learning algorithm. It analysis based on the statistics data for predication. The outcome of dependent variable is binary value. This value is used for evaluating the probability success.

***5.* KNN**

K-Nearest Neighbor (KNN) is a non-parametric and lazy supervised learning algorithm . This algorithm is used to have all cases and categorize new cases grounded on same measure. A KNN types an example to the class, which appears to be most determinedly among its k close by neighbors. k is a requirement for calibrating the characterization work.[4]

D. Experimental Result

The table 1 shows the experimental results for various algorithms which are used to predict the heart disease.

Algorithm Used	Accuracy
Decision Tree	72
SVM	68
Navie Baye's	57
Logistic Regression	68
KNN	63

Tabel 1. Accuracy Results.

III. Conclusion

Heart disease is one of the most important predictions for human beings because the death ratio is higher than other diseases all over the world. The dataset is used as an analysis parameter for training and testing purposes. Various algorithms are used to provide tremendously value in prognostic heart disease, which give more death rates throughout the world. As increasingly more work is being done in the field of Machine Learning, soon there might be new techniques to make Machine Learning more supportive in the field of medical services. The algorithms utilized in this trial have performed truly well utilizing the accessible properties. The conclusion can be finally drawn that machine learning is able to reduce the damage done to a person physically and mentally, by predicting heart disease.

IV. Reference

[1] Y. Khourdifi and M. Bahaj, "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization," *Int. J. Intell. Eng. Syst.*, vol. 12, no. 1, pp. 242–252, 2019, doi: 10.22266/ijies2019.0228.24.

[2] F. Fathieh *et al.*, "Predicting Cardiac Disease from Interactions of Simultaneously-Acquired Hemodynamic and Cardiac Signals," *Comput. Methods Programs Biomed.*, vol. 202, p. 105970, 2021, doi: 10.1016/j.cmpb.2021.105970.

[3] K. A. Syed Umar Amin Dr. Rizwan Beg, "Data Mining in Clinical Decision Support Systems for Diagnosis, Prediction and Treatment of Heart Disease," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 2, no. 1, pp. 218–223, 2013, [Online]. Available: http://uvic.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwY2BQSDFPM0xMMkk2SzNIS0mxSEoGnYCXYm6RZJFokmRunIKydQypNHcTYmBKzRNlcHNzDXH20AUtD4svgJy5EA86BRksAFsvFg_sf5tZJieZWZhaAlu5pqYpZmYpRuZJacAq2TgpzTDFQIyBBdiDTgUAlpgm7w.

[4] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Comput. Appl.*, vol. 29, no. 10, pp. 685–693, 2018, doi: 10.1007/s00521-016-2604-1.

[5] J. Soni, U. Ansari, and D. Sharma, "Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers," vol. 3, no. 6, pp. 2385–2392, 2011.

[6] S. Nikhar and A. M. Karandikar, "Prediction of Heart Disease Using Machine Learning Algorithms," *Int. J. Adv. Eng. Manag. Sci.*, vol. 2, no. 6, 2016, [Online]. Available: www.ijaems.com.

[7] S. Grampurohit and C. Sagarnal, "Disease prediction using machine learning algorithms," *2020 Int. Conf. Emerg. Technol. INCET 2020*, pp. 452–457, 2020, doi: 10.1109/INCET49848.2020.9154130.

[8] Purushottam, K. Saxena, and R. Sharma, "Efficient Heart Disease Prediction System," *Procedia Comput. Sci.*, vol. 85, pp. 962–969, 2016, doi: 10.1016/j.procs.2016.05.288.

[9] P. Perumal and P. T. Priyanka, "Supervised Heart Attack Prediction Using," vol. 7, no. 19, pp. 8089–8095, 2020.

[10] G. Parthiban, A. Rajesh, S. K. Srivatsa, and S. Professor, "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method," 2011.