Machine Learning Based Immense Investigation of Heart Disease

[1]Premananda Sahu, [2]Gandhari Chetan reddy, [3]Srikanth Thimmisetty,[4]Shaik Galib Shahid , [5]Akash Kumar

School of computer science engineering, lovely professional university

[1]premananda.29813@lpu.co.in,[2]chetangandhari@gmail.com,[3]thimmisettysrikanth@gmail.com,
[4]galibshahid714@gmail.com,[5]ak2330755@gmail.com

Abstract—**One of the leading causes of death is heart disease, often known as cardiovascular disease (CVD). all over the place. Approximately 17.9 million individuals worldwide pass away each year from heart disease, based on World Health Organization studies. According to the American heart association, heart disease causes 1 in 4 fatalities each year. The American Heart Association (AHA) estimates that the cost of heart disease to the US economy in terms of healthcare costs and lost productivity is $351.2 billion. In India, heart disease affects 2.8 million individuals annually and is estimated to cost the Indian economy $2.1 trillion between 2012 and 2030, according to a report by the American College of Cardiology.In this project, we suggested a web-based health recommendation system that can determine whether a person needs to see a doctor. On the Cleveland UCI dataset, the project compares various machine learning techniques, including LR, NB, KNN, and RF.The web-based recommendation system employs the most effective algorithm**

Keywords—cardiovascular disease, KNN, UCI dataset, HRS (Health recommendation system).

I. Introduction

Heart disease is the umbrella term for a variety of ailments which impact the functioning of the heart, such arrhythmias, heart failure, and coronary artery disease. According to up-to-date WHO statistics, 20.5 million people worldwide suffer from fatal heart conditions. Which causes a third of all deaths worldwide. Men who are middle-aged are more likely than women to develop heart disease.

Heart disease is caused by a heart that is unable to carry enough blood across the body. Heart disease risk factors include age, sex, smoking habits, obesity, alcohol use, and lack of physical activity. Heart disease indications comprise a chaotic heartbeat, a cold or sweet experiencing, nausea, discomfort in the chest, impulsive dizziness, and swollen ankles or feet. The patient's chance of survival will increase if the indications are correctly predicted. Right now, any healthcare choices are rendered by doctors just on a patient's medical history, today symptoms, or an electrocardiogram (ECG or EKG). And with a 67% accuracy velocity, doctors can predictions.

A method to diagnose CVD is angiography nevertheless it calls for an advanced level of expertise and can have undesirable consequences if utilised for evaluating blood vessels. This is one the explanations why more academics are turning to automated solutions lately. These techniques allow for the use of machine learning forecasting models like KNN, SVM, NB, AB, RF, and k-means cluster.

In recent times, there has been an increase in the amount of medical data generated, leading to unstructured and redundant data. Machine learning techniques have shown potential in predicting heart disease by using predictive models. However, to ensure the accuracy and efficiency of these models, it is necessary to pre-process the available data to extract relevant features and reduce the training time of the algorithm. Several studies have been conducted using various datasets, such as the Cleveland dataset, which have demonstrated promising results in terms of predicting heart disease accurately using machine learning algorithms. For instance, in one study, the J48 and LR models showed an accuracy of 56.8% and 55.8%, respectively, using the Cleveland dataset [1]

The suggested RS employs the highly accurate multi categorization algorithm (MOA) to choose the most pertinent characteristic accountable for the aetiology of heart disease. Additionally, it has attempted to provide forecasting models for precise illness forecasting. After fine-tuning the parameters and training, we utilize the KNN on the dataset gathered to create the predicting presence of heart disease or not. The average recall rate and accuracy to accurately anticipate illness prevalence were released.

II. RECENT STUDIES

An automated method that raises life quality and simplifies the work of medical professionals was proposed by Syed etal [2]. The system uses feature subset approach selection with performance improvement as its goal. Three algorithms—the mean-Fisher score selection method, the reverse feature selection method as well as the forward feature selection algorithm —are employed in each subset.

Each feature based on each of their Fisher scores, and the dimension and Matthews correlation coefficient score are utilised to select the feature subset.When the feature subset's reduced dimension is sent using an RBF kernel-based SVM, for example the set of features is split into two binary types: heart disease is present or not. The system's accuracy is 81.19%, 84.52%, 92.68%, and 82.7% for the Cleveland, Hungarian, Switzerland, and SPECTF datasets, respectively.

Cengiz [3] proposed research to study all the best possible classification methods for detecting any heart diseases. For the study researchers took 18 machine learning methods. These approaches are broken down into six distinct categories, and three selections of features are made. The eighteentechniques used are LR, J-48, NB, KNN, RS, MLP, SVM,SVM-RBFkernal, RF, nbtree, rbf net, FR-NN, fuzzyNN, N-N, GP. In the paper the techniques for selecting the appropriate features are not selecting any features, CFS, FRS and chi-square. In accordance to the paper SVM gave the most accuracy for no feature selection 85.2%, and for CFS selection highest accuracy is given by NB with 84.9%.

Research was proposed by Daniel et al [4]. to reduce the number of tests necessary to determine if a patient has cardiac disease or not. After uncovering hidden patterns utilizing data mining and data research approaches, in order to determine whether cardiac disease existed, the study used machine learning methods such LLR,DTclassification, and GNB model. All of the information has been processed beforehand, for as it eliminates value duplication from the rows, before stats are collected from the repository maintained by UCI. deletion, or adding details to the missing sections, in addition to deleting outlines that signify values that are in excess of three standard deviations out of the mean. Next, the data are transformed and reduced, after that all three models are trained on the data, and K-fold cross validation is used to determine the outputs.

A diagnostic strategy for CHD employing MLpractices such multi classifier system was proposed by Bayu et al[5]. The Cleveland, Hungarian, and Z-Alizadeh Sani datasets are where the data came from. A two-tier ensemble structure exists in the system, where one ensemble categorization is used by another ensemble. Extreme gradient boosting, gradient boosting machine, and random forest are three ensemble learners. —are used to design an architecture. In the initial stage of the suggested system, we choose the features we need by removing irrelevant characteristics using CFS (correlation-based feature selection), and then we apply optimisation techniques like PSO (particle swarm optimisation) to improve search.Following the feature selection process, classification techniques are utilised to create a two-tier ensemble utilising three separate classifiers. With an accuracy of 83.905%, the 20 particle PSO prediction is the best one made using the Z-Alizadeh Sani dataset.

Hui et al. [6] proposed a classification system for the Cleveland heart disease dataset using a three-attribute evaluation approach. The dataset was pre-processed by cleaning the data, and then 10 different machine learning classifiers were employed, including naïve Bayes, logistic regression, sequential minimal optimizer, IBk, and adaBoostM1. A ten-fold cross-validation testing method was employed to determine the algorithm's performance. The technique via SMO has the greatest reliability, 85.1%. with a MAE of 0.15, precision of 0.855, and Harmonic mean of 0.852. In contrast, the j-Rip algorithm had the lowest accuracy of 74.59%.

III. METHODOLOGY

Fig.1 Project flow of proposed work

In the Project Flow we tried to demonstrate how Project works.

Data from the Kaggle UCI heart disorders data set are collected as part of the project's initial stage. age, sex, cp, testbps, chol, fbs, restecg,thalach, exang, old peak, slope, ca, thal, target are the 14 attributes of the multivariant data set that are displayed in Table 1.

TABLE 1. ATTRIBUTE/PROPERTIES DESCRIPTION TABLE

Attribute	Description	Attribute Type
Patient age	Age of the patient	Integer
Sex	Male or female	categories
Chest pain-cp	Chest pain type	categories
testbps	Blood pressure in millimetres when at rest	Integer
chol	mg/dL of total cholesterol in the blood	Integer
fbs	>120 mg/dL is the lowest blood sugar level.	categories
restecg	electrocardiographic readings when at rest	categories
thalach	reached highest heart rate	Integer
exang	Pressure brought on by exercise	categories
oldpeak	Compared to a state of rest, there is ST depression induced by exercise.	Integer
slope	The ST segment angle during peak	categories

	exercise	
ca	cinefluorography-visible vessel count	categories
thal	a cardiac condition	categories
Target	Predictive quality	categories

Table 2 lists properties like "Age," "testbts," "chol," and "thalach," that are defined by the integer type and consist of values that fall between certain ranges of numbers. Other properties possess categories, which implies that the values given to individuals depend on the category to which they were in. Male = 1 and female = 0 are the two classes for the attribute known as "sex." A patient's chest discomfort is of the Cp kind. The property has four classifications. Asymptomatic angina is indicated by the numbers 1, 2, and 4. The fasting blood sugar is fbs. Depending on whether the sugar concentration is more than 120 mg/dL determines how the characteristic is divided. If the sugar concentration is greater than 120 the attribute is categorized in to class 1, if not the attribute is categorized in to class 0.The attribute for categorizing electrocardiograph image while the patient is at rest is restecg, based on the observation by the doctor the attribute get a value 0,1,2- if the patient has normal electrocardiograph result the attribute is give class 0, if the ST and T waves is showing any abnormality then the attribute is assigned class value 1. If the electrocardiograph shows thickening of heart pumping chambers the attribute is given class value of 2. The chest pain caused because the less flow of blood to the heart is categorized as exang the category is divided in to 2 classes- 0 if there is no lwo blood pressure found in the heart and 1 if the blood pressure is found. The slope of the ST wave in electrocardiograph is categorized as slope and divided in to 3 categories. The second feature, ca, which is the number of vessels that can be seen via fluoroscopy, ranges from 0 to 3. The heart shape is determined by the thal attribute, which has three classes: 3 for no change, 6 for defect that is fixed, and 7 for defect that is reversible. The final attribute, target, has two classes that indicate whether there is a risk of developing a heart condition: 0 for no risk and 1-4 for a range of risks.

The collection of histograms in the Figure 2 are the histograms of attribute that are present in the dataset. Using the data present in the histogram we can find the max, min, mean and std dev values of numeric attributes and number of times classes repeated in the categoric attribute. Age which is a numeric attribute has values lies between late 20’s and mid 70’s with mean 54 and std dev of 9. testbps is another numeric attribute where the min and max values lies between 90 and 200 with mean of 131 and std dev of 17.Chol's ranges from 126 to 564, with anaverage of 246.69 and a std dev of 51.777. Thalach has anaverage of 149.607 and a std dev of 22.875, with a minimum of 71 and a high of 202. Openpeak's ranges from 0 to 6.2, with an average of 1.04 and a std dev of 1.

Sex is a categorical attribute which has female and male classes with 97 and 202 frequencies of respective classes. Cp is the next categorical attribute which is divided into 4 classes each classes having frequencies of 23, 50, 86 and 144. The sugar concentration is given to a categorical attribute fbs which has 2 classes with frequencies 258 and remaining 45 belongs to another class. The electrocardiograph of the patient at rest is give to a categorical attribute restecg which has 3 classes with frequencies 151, 4, 148. Low blood pressure during exercise is given to a categorical attribute with 2 classes frequencies of each class are 204, 99. The slope of ST graph in electrocardiograph is given to a categorical attribute slope with frequencies 142, 140, 21.

Similarly, the max-heart rate is given to attribute of categorical type which has 3 classes of frequencies 166, 18, 117 and their classes are 3, 6, 7 respectively. Target is the last attribute which is also a categorical type which has two classes with frequencies of 164, 139.

the above data is acquired from University of California repository. In the preprocessing the data is checked for quality such weather the attribute has the value null or not null. then the data is visualized to show the gender of the patient. Histogram of all the attributes is created to give visualization to the data. later the data is split for training and testing with .8 for training and .2 for testing after that the data set is trained with LR, KNN, NB, DT, RF classifiers which leads to accuracy of 77.05%, 94.79%, 78.69%, 67%, 80.33% respectively.

As KNN gave the highest accuracy, next step is to create a web site where an admin can create users and each users are given access to a form where the user enters the patient details like what kind of chest pain does the patient have, what is the blood pressure of the patient while on rest, cholesterol of the patient in mg, does the patient sugar concentration is more than 120 mg or not, and other details like electrocardiograph and heart rate etc. then the user will press predict then using kNN classifier the website will return weather the patient should consult the doctor for IN person checkup or not.

Logistic regression-

The supervised machine learning technique logistic regression provides discrete values for categorical dependent variables. Instead of providing a true or false answer, it provides a probability between 0 and 1.

Logistic regression comes in three different flavours (1). binary that just has pass or fail as the dependent variables. (2) A multinomial with three or more dependent variables that are not ordered. three or more ordered dependent variables are present in an ordinal, in (3).

P= 1/ (1 + e^(-a))

In this case, a=c0 + c1*y1 + c2*y2 +.. + cn*yn.

K-nearest neighbor-

A supervised learning machine learning algorithm is K-nearest neighbor. KNN determines if new data and existing data are comparable and assigns the new data to the category that is most like the category of existing data. The lazy learner algorithm is the simplest algorithm. KNN is a non-parametric method, therefore it makes no assumptions about the underlying data.

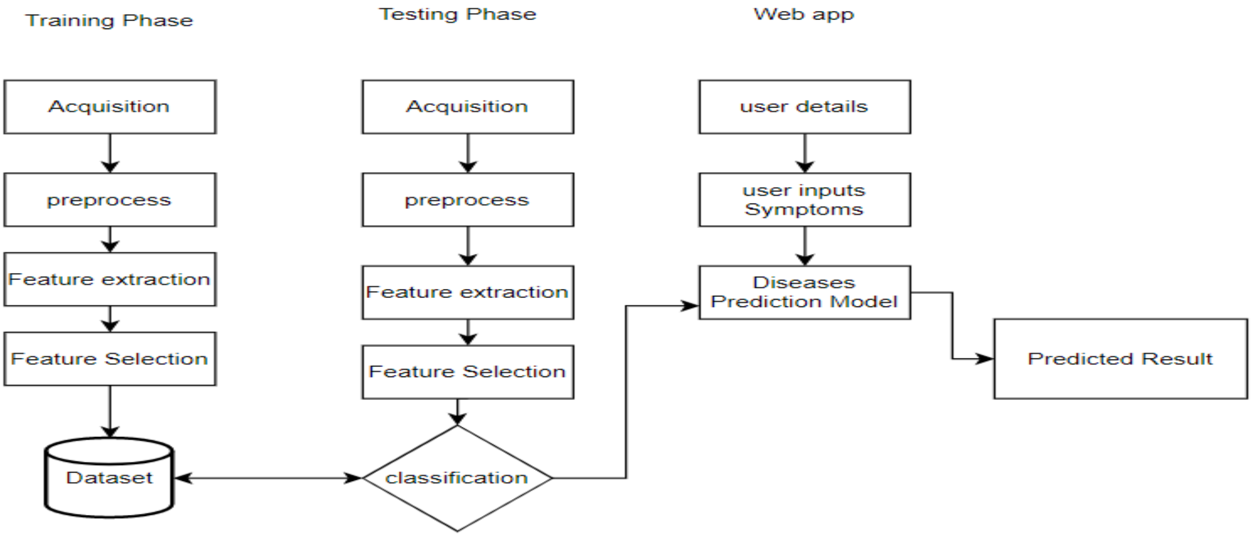


Figure 1- Architectural diagram of proposed work

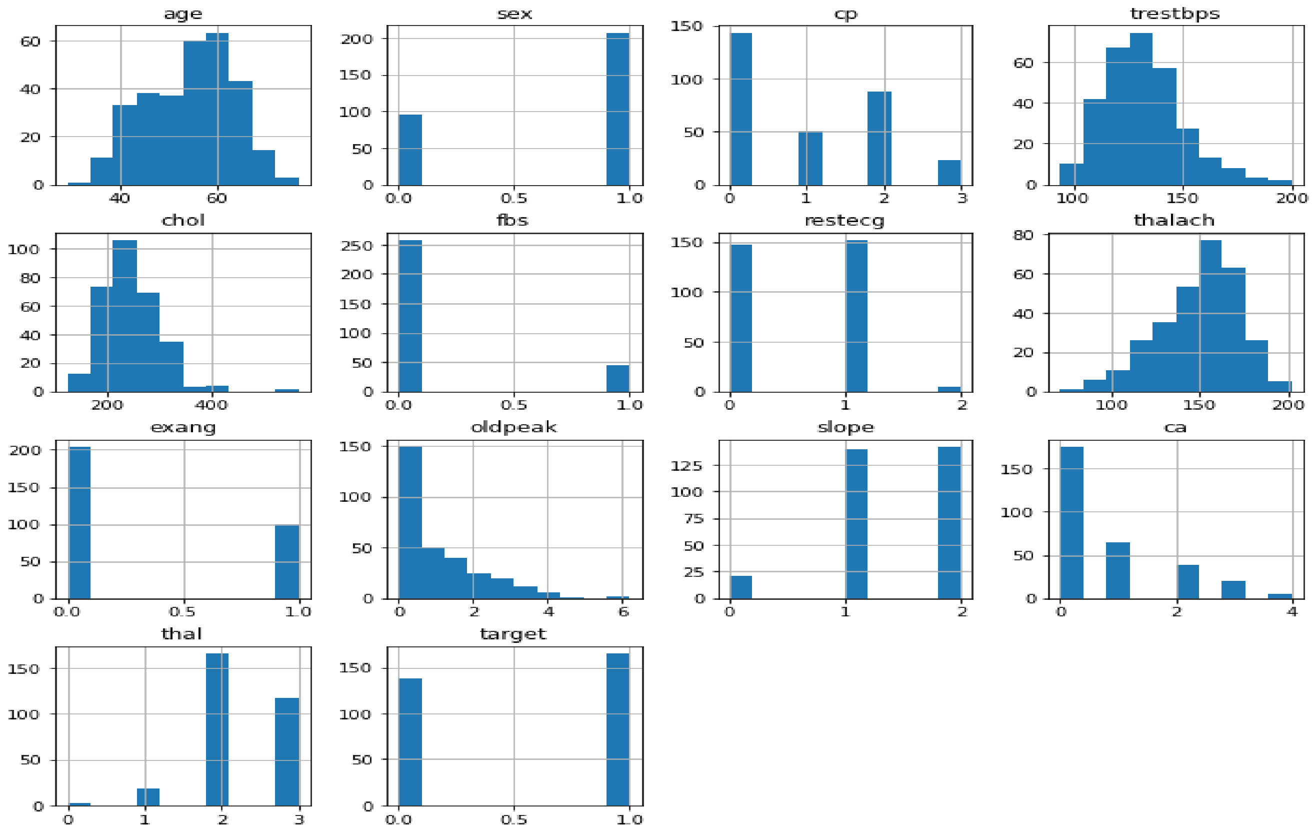


Figure 2- Histogram of Attributes in data set

Naïve bayes-

A supervised machine learning technique called naive bayes employs the Bayes theorem under the naive presumption that each pair of feature pairs is independent of the others. Naive Bayes predicts the outcome based on the likelihood of the input.

Decision Tree-

A supervised machine learning technique that can do regression in addition to classification is the decision tree [7]. The framework is a tree, with nodes denoting dataset characteristics and edges denoting selection criteria. Therefore, every potential solution to the problem is shown in the graphical depiction.

Random forest-

Random forest, which is based on the idea of ensemble learning, is used for both classification and regression. Decision trees are included in random forests, which use diverse subsets of datasets and then average them to increase accuracy. When there are vast amounts of missing data, random forest is utilised to maintain excellent accuracy. It also trains very quickly.

IV. RESULTS

The classification results for the algorithms for logistic regression, nave bayes, k-nearest neighbors, decision trees, and random forests are shown in the tables below.

Logistic regression-

TABLE 2- PERFORMANCE EVALUATION REPORT OF LOGISTIC REGRESSION

	positive predictive value	True positive rate	Harmonic mean	Frequency of occurrence
Female	0.76	0.70	0.73	27
male	0.78	0.82	0.80	34
Classification rate			0.77	61
macro-averaging	0.77	0.76	0.77	61
Weighted-averaging	0.77	0.77	0.77	61

The classification report is shown above is for Logistic Regression. For Female the Positive Predictive value is 76%, True Positive rate is 70%, Harmonic mean is 73% and Frequency of occurrence is 27. For male the Positive Predictive value is 78%, True Positive rate is 82%, Harmonic mean is 80%, Frequency of occurrence is 34. LR boasts an overall accuracy rate that is 78.69%.

Naïve bayes-

TABLE 3- PERFORMANCE EVALUATION REPORT OF NAÏVE BAYES

	Positive predictive value	True positive rate	Harmonic mean	Frequency of occurrence
Female	0.82	0.67	0.73	27
Male	0.77	0.88	0.82	34
Classification rate			0.79	61
macro-averaging	0.79	0.77	0.78	61
Weighted-averaging	0.79	0.79	0.79	61

The classification report is shown above is for Naïve Bayes. For Female the Positive Predictive value is 82%, True Positive rate is 67% Harmonic mean is 73% and Frequency of occurrence is 27. For male the Positive Predictive value is 77%, True Positive rate is 88%, Harmonic mean is 82%, Frequency of occurrence is 34. NB boasts an overall accuracy rate that is 78.70%.

KNN-

TABLE 4- PERFORMANCE EVALUATION REPORT OF KNN.

	Positive predictive value	True positive rate	Harmonic mean	Frequency of occurrenc

				e
Female	0.93	0.93	0.94	27
male	0.90	0.91	0.93	34
Classification rate			0.94	61
macro-averaging	0.94	0.95	0.94	61
Weighted-averaging	0.93	0.96	0.95	61

The classification report is shown above is for K-nearest neighbor. For Female the Positive Predictive value is 93%, True Positive rate is 93%, Harmonic mean is 94% and Frequency of occurrence is 27. For male the Positive Predictive value is 90%, True Positive rate is 91%, Harmonic mean is 93%, Frequency of occurrence is 34. KNN boasts an overall accuracy rate that is 94.79%.

Decision Tree-

TABLE 5- PERFORMANCE EVALUATION REPORT OF DECISION TREE

	Positive predict value	True positive rate	Harmonic mean	Frequency of occurrence
Female	0.68	0.53	0.59	27
Male	0.69	0.80	0.74	34
Classification rate			0.68	61
macro-averaging	0.68	0.67	0.67	61
Weighted-averaging	0.68	0.68	0.67	61

The classification report is shown above is for Random Forest. For Female the Positive Predictive value is 68%, True Positive rate is 53%, Harmonic mean is 59% and Frequency of occurrence is 27. For male the Positive Predictive value is 69%, True Positive rate is 80%, Harmonic mean is 74%, Frequency of occurrence is 34. Decision Tree boasts an overall accuracy rate that is 67.3%.

Random forest-

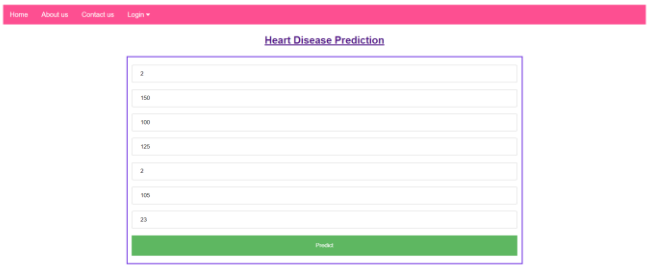
TABLE 6 – PERFORMANCE EVALUATION REPORT OF RANDOM FOREST

	Positive predictive value	True positive rate	Harmonic mean	Frequency of occurrence
Female	0.86	0.67	0.75	27
Male	0.78	0.91	0.84	34
Classification rate			0.81	61
macro-averaging	0.83	0.80	0.80	61
Weighted-averaging	0.82	0.81	0.81	61

The classification report is shown above is for Random Forest. For Female the Positive Predictive value is 86%, True Positive rate is 67%, Harmonic mean is 75% and Frequency of occurrence is 27. For male the Positive Predictive value is 78%, True Positive rate is 91%, Harmonic mean is 84%, Frequency of occurrence is 34. Random Forest boasts an overall accuracy rate that is 80.4%.

The aforesaid kNNhas the highest accuracy and is used for web-apps. The user will be submitting their details in a web-based application that is demonstrated in fig.3. there are sevensubmit fields "chest pain type", "resting blood pressure", "serum cholesterol in mg/dL", "fasting blood sugar", "resting electrocardiographic results", "max heart rate achieves", "exercise induced angina a". the patient data was entered, As shown in Fig. 4, the system will let the patient figure out whether they have heart disease or not and advise them to see a doctor.

Figure 3- HRS application using web-based technology



Heart Disease Prediction

2

150

100

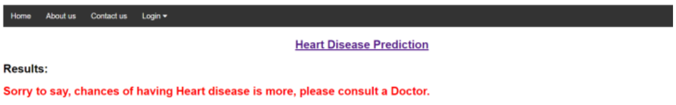
125

2

105

23

Figure- 4



Heart Disease Prediction

Results:

Sorry to say, chances of having Heart disease is more, please consult a Doctor.

V. Conclusion

A notable advancement in the healthcare is research in ML based recommendation system for assessing the chances of heart diseases. The system analyses patient data using machine learning techniques to forecast the likelihood of recurrent cardiac issues. With more knowledge, this can assist medical professionals in choosing the best course of treatment for a patient. More precise predictions are achieved when machine learning algorithms are applied successfully to predict illness. The model distinguishes between several forms of cognitive impairment in addition to predicting the patient's illness.

REFERENCES

[1]Patel, J., 2015. Prof. Tejal Upadhyay, Dr. Samir Patel,“. Heart disease prediction using Machine learning and Data Mining Technique, 7(1Sept), p.2016.

[2]Saqlain, Syed Muhammad, Muhammad Sher, Faiz Ali Shah, Imran Khan, Muhammad Usman Ashraf, Muhammad Awais, and Anwar Ghani. "Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines." Knowledge and Information Systems 58 (2019): 139-167.

[3]Gazeloglu, C. (2020) “Prediction of heart disease by classifying with feature selection and machine learning methods ”, Progress in Nutrition, 22(2), pp. 660–670. doi: 10.23751/pn.v22i2.9830.

[4]Ananey-Obiri D, Sarku E. Predicting the presence of heart diseases using comparative data mining and machine learning algorithms. International Journal of Computer Applications. 2020;176(11):17-21.

[5]Bayu Adhi Tama, Sun Im, Seungchul Lee, "Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier Classifier Ensemble", BioMed Research International, vol. 2020, Article ID 9816142, 10 pages, 2020.

[6]Reddy, Karna Vishnu Vardhana, IrraivanElamvazuthi, Azrina Abd Aziz, SivajothiParamasivam, Hui Na Chua, and S. Pranavanand. "Heart disease risk prediction using machine learning classifiers with attribute evaluators." Applied Sciences 11, no. 18 (2021): 8352.

[7]B. Charbuty and A. Abdulazeez, “Classification Based on Decision Tree Algorithm for Machine Learning”, JASTT, vol. 2, no. 01, pp. 20 - 28, Mar. 2021.