IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

Heart Diseases Prediction System using ML

Prof. Madhavi Tota[1], Manthan Moon[2], Pranit Nagrale[3], Akshay Pandav[4], Gunjan Das[5]
Guide, Department of Information Technology[1]
Students, Department of Information Technology[1,2,3,4]
Rajiv Gandhi College of Engineering, Research and Technology, Chandrapur, Maharashtra, India
Dr. Babasaheb Ambedkar Technical University, Lenore

Abstract: *Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health care industry. This project aims to predict future heart disease by analysing data of patients which classifies whether they have heart disease or not using machine learning algorithm. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally ana lysing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease.*

Keywords: Machine Learning

I. INTRODUCTION

According to the World Health Organization, every year 12 million deaths occur worldwide due to heart disease. Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of data analysis. The load of cardiovascular disease is rapidly increasing all over the world from the past few years[1]. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high turn reduces the complications.

Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health care industry. This project aims to predict future heart disease by analysing data of patients which classifies whether they have heart disease or not using machine learning algorithm[2]. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally ana lysing to extract the desired data we can say that this technique can be very well adapted to do the prediction of hear t disease[3].

II. WORKING AND EXISTING SYSTEM

The existing system modules generates comprehensive report by implementing the strong prediction algorithm. In this project the input details are obtained from the patient and the doctor. Then from the doctor inputs, using ai algorithms heart disease is analysed. Now, the obtained result is compared with the result of existing models with in the same domain and found to be improved[2]. The main aims of the existing system to compare and check the before patient whose having disease outputs and new patient disease and determine future possibilities of the heart disease to a particular patient By Implementing the above-mentioned model we will get the goal of developing a system with increased rate of accuracy of estimating the new patient getting heart attack percentage. The data of heart disease patients collected from the UCI laboratory is used to discover patterns with K Neighbours Classifier, Support Vector Classifier, decision Tree Classifier, Random Forest Classifier. The results are compared for performance and accuracy with these AI algorithms. The model which is proposed for Heart Disease Prediction System is invented for using different algorithms of AI and approach. But by using all the existing systems the accuracy is very less[5].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

Authors: Zaibunnisa L. H. Malik, Momin Fatema, Nikam Pooja, Gawandar Ankita Heart Disease Prediction using Artificial Intelligence – IJERT

Now a day's human beings are so busy in their existence to reap what they need and earn that they overlook to take care of their fitness. Because of this, there's alternate inside the food which they devour, their life- style changes[1]. They are greater tensed and in very much strain to earn money so this results in blood strain, diabetes and diverse different illnesses at young age. All those reasons result in negligence in their fitness which will increase the chances of heart disorder. Heart is the maximum crucial organ of the human body and if it is affected then it additionally affects the alternative most important organs of the frame[3].

Clinical choices are frequently made primarily based on physicians' intuition and experience in preference to on the knowledge rich facts hidden in the database[1]. This practice ends in unwanted biases, errors and excessive medical expenses which affects the satisfactory of carrier supplied to patients. HD diagnosis traditionally by using medical history of patient. However, the diagnosis results are not accurately diagnosis HD. Furthermore, these methods are not reliable in terms of accuracy and computation[2]. There are a number of publications that propose different techniques for the extraction of features from the heart sounds and classify them using neural networks. In the late 80's Mohamed and Raafat developed a mathematical model to describe the heart sounds and murmurs by a finite number of parameters. In this case, features were extracted based on fourth order linear prediction of the cardiac cycle frames, where classification was carried out based on the minimum distance between the features of the measured pattern and the reference patterns. Patil and Kumaraswamy proposed an intelligent heart attack prediction system based on Data Mining and Artificial Neural Network[15].

In this method, the parameters vital to the heart attack are computed by using K-means clustering algorithm to the available data. These frequent patterns are mined from the data, with the aid of the Maximal Frequent Itemset Algorithm (MAFIA)[4]. The patterns are then selected based on the computed significant weight age. Although the above study reported that this method is capable of predicting the heart attack using MAFIA algorithm, the prediction accuracy was not reported for the work[5]. Furthermore, this technique uses features corresponding to the behavioural habits of the subject, such as smoking and alcohol consumption, instead of feature characteristics of the heart sound signal itself. In this project we have implemented ML algorithms such as:

1. K Neighbours Classifier
2. Support Vector Classifier
3. Decision Tree Classifier
4. Random Forest Classifier

Which can predict heart disease and we are first taking input from doctor about heart related information that is smoking, cholesterol, high blood pressure etc and then our system will predict the heart disease from given algorithms and will define that which algorithm is best for prediction of Heart disease.

Author.: - Pabitra Kumar Bhunia, Arijit Debnath, Poulami Mondal, Monalisa D E, Kankana Ganguly, Pranati Rakshit (Department of Computer Science and Engineering JIS College of Engineering Kalyani, Nadia, West Bengal, India) Numerous studies have been done that have focused on the diagnosis of heart disease. They have applied different data mining techniques for diagnosis & achieved different probabilities for different methods.

This system evaluates those parameters using the data mining classification technique. The datasets are evaluated in python using two main Machine Learning Algorithms: The decision Tree Algorithm and the Naive Bayes Algorithm which shows the best algorithm between these two in terms of the accuracy level of heart disease[5] .

Aditi Gavhane - Predicted heart attack for early diagnosis to reduce the count of deaths. For this problem Machine Learning plays a major role in this paper. This prediction takes people from the danger zone of their life. In this paper, we use the KNN algorithm and Random Forest algorithm to predict the heart attack in advance [3].

Senthil Kumar - Introduced a prediction model with different combinations of features, and several known classification techniques. It produced an enhanced performance level with an accuracy level of 88.7% through the prediction model for heart disease with Hybrid Random Forest with Linear Model (HRFM)[11].

Himanshu Sharma - Stated and proved that machine learning algorithms and deep learning opens new door opportunities for precise prediction of a heart attack. Paper provides a lot of information about state of art methods in

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

Machine learning and deep learning. An analytical comparison has been provided to help new researchers working in this field[7] .

M. Nikhil Kumar - W orked with 8 algorithms including Decision Tree, J48 algorithm, Logistic model tree algorithm, Random Forest algorithm, NaÃ¯ve Bayes, KNN, Support Vector Machine, Nearest Neighbor to predict heart diseases. The accuracy of the prediction level is high when using more attributes[8].

Amandeep Kaur- Stated that Data mining is an important stage of the KDD process that can be used for disease management, diagnosis, and prediction in healthcare organizations. This paper discusses reviews on different methods and approaches in data mining that have been used to predict heart disease[5].

Pahulpreet Singh Kohli developed an Enhanced New Dynamic Data Processing (ENDDP) Algorithm to predict the early stages of heart disease. The results prove the performance of the proposed system

Author - Senthilkumar Mohan, ChandrasegarThirumalai, Gautam Srivastava .Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques, Digital Object Identifier, 10.1109/ACCESS.2019.2923707, IEEE Access, VOLUME 7, 2019

There is number of works has been done related to disease prediction systems using different machine learning algorithms in medical Centres.

Senthil Kumar Mohan - Proposed Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques in which strategy that objective is to finding critical includes by applying Machine Learning bringing about improving the exactness in the expectation of cardiovascular malady. The expectation model is created with various blends of highlights and a few known arrangement strategies. We produce an improved exhibition level with a precision level of 88.7% through the prediction model for heart disease with hybrid random forest with a linear model (HRFLM) they likewise educated about Diverse data mining approaches and expectation techniques, such as, KNN, LR, SVM, NN, and Vote have been fairly famous of late to distinguish and predict heart disease[10].

Sonam Nikhar - has built up the paper titled as Prediction of Heart Disease Using Machine Learning Algorithms by This exploration plans to give a point by point portrayal of NaÃ¯ve Bayes and decision tree classifier that are applied in our examination especially in the prediction of Heart Disease[4]. Some analysis has been led to think about the execution of prescient data mining strategy on the equivalent dataset, and the result uncovers that Decision Tree beats over Bayesian classification system[5].

Aditi Gavhane, GouthamiKokkula, Isha Pandya, Prof. Kailas Devadkar (PhD) - Prediction of Heart Disease Using Machine Learning, In this paper proposed system they used the neural network algorithm multi-layer perceptron (MLP) to train and test the dataset. In this algorithm there will be multiple layers like one for input, second for output and one or more layers are hidden layers between these two input and output layers. Each node in input layer is connected to output nodes through these hidden layers. This connection is assigned with some weights. There is another identity input called bias which is with weight b, which added to node to balance the perceptron. The connection between the nodes can be feedforwarded or feedback based on the requirement[13].

Abhay Kishore - developed Heart Attack Prediction Using Deep Learning in which This paper proposes a heart attack prediction system using Deep learning procedures, explicitly Recurrent Neural System to predict the probable prospects of heart related infections of the patient. Recurrent Neural Network is a very ground-breaking characterization calculation that utilizes Deep Learning approach in Artificial Neural Network[15]. The paper talks about in detail the significant modules of the framework alongside the related hypothesis. The proposed model deep learning and data mining to give the precise outcomes least blunders. This paper gives a bearing and point of reference for the advancement of another type of heart attack prediction platform and Prediction stage[5].

Lakshmana Rao - Machine Learning Techniques for Heart Disease Prediction in which the contributing elements for heart disease are more (circulatory strain, diabetes, current smoker, high cholesterol, etc..). So, it is difficult to distinguish heart disease. Different systems in data mining and neural systems have been utilized to discover the seriousness of heart disease among [17]. The idea of CHD ailment is bewildering, in addition, in this manner, the disease must be dealt with warily. Not doing early identification, may impact the heart or cause sudden passing. The perspective of therapeutic science furthermore, data burrowing is used for finding various sorts of metabolic machine

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

learning a procedure that causes the framework to gain from past information tests, models without being expressly customized[16]. Machine learning makes rationale dependent on chronicled information.

Mr. SanthanaKrishnan.J and Dr.Geetha.S, - Prediction of heart disease using machine learning algorithm This Paper predicts heart disease for Male Patient using Classification Techniques. The detailed information about Coronary Heart diseases such as its Facts, Common Types, and Risk Factors has been explained in this paper. The Data Mining tool used is WEKA (Waikato Environment for Knowledge Analysis), a good Data Mining Tool for Bioinformatics Fields. The all three available Interface in WEKA is used here; Naive Bayes, Artificial Neural Networks and Decision Tree are Main Data Mining Techniques and through this techniques heart disease is predicted in this System[5]. The main Methodology used for prediction is Decision Trees like CART, C4.5, CHAID, J48, ID3 Algorithms, and Naive Bayes Techniques.

Avinash Golande - Proposed Heart Disease Prediction Using Effective Machine Learning Techniques in which Specialists utilize a few data mining strategies that are available to support the authorities or doctors distinguish the heart disease[16]. Usually utilized methodology utilized are decision tree, k- closest and NaÃ¯ve Bayes. Other unique characterization-based strategies utilized are packing calculation, Part thickness, consecutive negligible streamlining and neural systems, straight Kernel self- arranging guide and SVM (Bolster Vector Machine)[17]. The following area obviously gives subtleties of systems that were utilized in the examination.

V.V. Ramalingam - Proposed Heart disease prediction using machine learning techniques in which Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data. Many researchers, in recent times, have been using several machine learning techniques to help the health care industry and the professionals in the diagnosis of heart related diseases. This paper presents a survey of various models based on such algorithms and techniques and analyse their performance. Models based on supervised learning algorithms such as Support Vector Machines (SVM), K- Nearest Neighbour (KNN), NaÃ¯ve Bayes, Decision Trees (DT), Random Forest (RF) and ensemble models are found very popular among the researchers and systems have been applied to different clinical datasets to robotize the investigation of huge and complex information[3]. Numerous scientists, as of late, have been utilizing a few Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data. Many researchers, in recent times, have been using several machine learning techniques to help the health care industry and the professionals in the diagnosis of heart related diseases. This paper presents a survey of various models based on such algorithms and techniques and analyze their performance. Models based on supervised learning algorithms such as Support Vector Machines (SVM), K- Nearest Neighbour (KNN), NaÃ¯ve Bayes, Decision Trees (DT), Random Forest (RF) and ensemble models are found very popular among the researchers. strategies to enable the wellbeing to mind industry and the experts in the analysis of heart related sicknesses[9]. This paper presents a review of different models dependent on such calculations and methods and analyze their exhibition. Models in light of directed learning calculations, for example, Support Vector Machines (SVM), K- Nearest Neighbour (KNN), NaÃ¯ve Bayes, Decision Trees (DT), Random Forest (RF) and group models are discovered extremely well known among the scientists.

Author - Aditya Khamparia (Assistant Professor) at Babasaheb Bhimrao Ambedkar University (Central University), India

The correct prediction of heart disease can prevent life threats, and incorrect prediction can prove to be fatal at the same time. Heart disease describes a range of conditions that affect your heart. Today, cardiovascular diseases are the leading cause of death worldwide with 17.9 million deaths annually, as per the World Health Organization reports[6] . Various unhealthy activities are the reason for the increase in the risk of heart disease like high cholesterol, obesity, increase in triglycerides levels, hypertension, etc. Many studies have been performed and various machine learning models are used for doing the classification and prediction for the diagnosis of heart disease[8]. An automatic classifier for detecting congestive heart failure shows the patients at high risk and the patients at low risk

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

Pre-processing of the Dataset -The dataset does not have any null values. But many outliers needed to be handled properly, and also the dataset is not properly distributed. Two approaches were used. One without outliers and feature selection process and directly applying the data to the machine learning algorithms, and the results which were achieved were not promising. But after using the normal distribution of dataset for overcoming the overfitting problem and then applying Isolation Forest for the outlier's detection, the results achieved are quite promising[5].

Checking the Distribution of the Data - 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 overfit[16].

Checking the Skewness of the Data.- For checking the attribute values and determining the skewness of the data (the asymmetry of a distribution), many distribution plots are plotted so that some interpretation of the data can be seen. Different plots are shown, so an overview of the data could be analyzed. The distribution of age and sex, the distribution of chest pain and trestbps, the distribution of cholesterol and fasting blood, the distribution of ecg resting electrode and thalach, the distribution of exang and oldpeak, the distribution of slope and ca, and the distribution of thal and target all are analyzed and the conclusion By analyzing the distribution plots, it is visible that thal and fasting blood sugar is not uniformly distributed and they needed to be handled; otherwise, it will result in overfitting or underfitting of the data[12].

Checking Stats of the Normal Distribution of Data - Checking the features which are important for heart disease and not important for heart disease, respectively. Here the important factors show a different variation which means it is important. The conclusion which can be drawn from these statistical figures is that we can see a Gaussian distribution which is important for heart disease and no Gaussian distribution which is playing that much important role in heart disease [9].

Feature Selection. - For selecting the features and only choosing the important feature, the Lasso algorithm is used which is a part of embedded methods while performing feature selection. It shows better predictive accuracy than filter methods. It renders good feature subsets for the used algorithm. And then for selecting the selected features, select from the model which is a part of feature selection in the scikit-learn library[4].

Checking Duplicate Values in the Data - The duplicates should be tackled down safely or otherwise would affect the generalization of the model. ere might be a chance if duplicates are not dealt with properly; they might show up in the test dataset which is also in the training dataset[16].

Machine Learning Classifiers Proposed - The proposed approach was applied to the dataset in which firstly the dataset was properly analysed and then different machine learning algorithms consisting of linear model selection in which Logistic Regression was used[9]. For focusing on neighbor selection technique KNeighbors Classifier was used, then tree-based technique like Decision Tree Classifier was used, and then a very popular and most popular technique of ensemble methods Random Forest Classifier was used. Also, for checking the high dimensionality of the data and handling it, Support Vector Machine was used. Another approach which also works on ensemble method and Decision Tree method combination is XGBoost classifier[12] .

Author - B.Venkatalakshmi, and M.V Shivsankar, "Heart Disease Diagnosis Using Predictive Datamining", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 3, Special Issue 3, March2014.
Big data defines as large volumes of high velocity, complex, and shifting data that require advanced skills and technologies to set up the capture, storage, distribution, management and analysis of the information. Big data enclose such characteristics as variety, velocity and, with respect especially to healthcare, integrity. Current analytical techniques can be applied to the large amount of existing patient related health and medical data[16].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

To reach lower understanding of results, which then can be applied at the point of care. Ideally, unique and population data would inform each specialist and her /his patient while the decision making process and help Regulate the most proper treatment option for that particular patient.

Big data in healthcare refers to electronic health data sets so large and complex that it is difficult to manage with traditional or common data management methods and traditional software and/or hardware[1]. Some health care data are characterized by a need for timeliness; for example, data generated by wearable or implantable biometric sensors; blood pressure, or heart rate is often required to be collected and analyzed in real-time[5]. Data in healthcare can be categorized as follows

Clinical Data and Clinical Notes About 80% of this type data are unstructured documents, images and clinical or transcribed process.

Structured data (e.g., laboratory data, structured EMR/HER)

Unstructured data (e.g., post-op notes, diagnostic testing reports, patient discharge summaries, unstructured EMR/HER and medical images such as radiological images and X-ray images)

Semi-structured data (e.g., copy-paste from other structure source)

Author -Pooja Anbuselvan Student at Bangalore Institute of Technology Bengaluru, Karnataka, India.

Cardio-Vascular diseases are the primary cause of death worldwide over the past decade. According to the World Health Organization it is estimated that over 17.9 million deaths occur each year because of cardiovascular diseases and out of these deaths 80% is attributed to coronary artery disease and cerebral stroke [1]. Many habitual factors such as personal and professional habits and genetic predisposition accounts for heart disease The major challenge faced in the world of medical sciences today is the provision of quality service and efficient and accurate prediction. Learning (ML) which is subfield of data mining that deals with large scale well-formatted dataset efficiently[5]. In the medical field, machine learning can be used for diagnosis, detection and prediction of various diseases.

III. OBJECTIVE OF RESEARCH

- Classification prediction model to just about all of the foremost extremely cited distance measures within the connected on heart condition datasets.
- Nearest Neighbor classifiers are the most effective method and take into account alternative classifiers, as well as neural networks and J.48 algorithm.
- To improve analysis study area by increasing our search area to incorporate deletion.
- Neural network approach is captures the optimization values and uses these values to represent the statistic measurement.
- PSO optimization selection model to detect the key modification points in anytime series, and uses these points to represent the whole statistic.
- NN classification approach is to interrupt a variable statistic instance into multiple univariate-time series data then every time series is processed individually into disjoint segments and also the aggregate distance is generated.

IV. PERFORMANCES METRICS ANALYSIS

It describes an evaluation metrics for heart disease prediction model. The table contains Mean Absolute error, Root Relative square Error, Root Relative Square Error and Accuracy values of SVM, KNN, RF, J.48 and MLP classification algorithm.

Conclusion

The proposed technique is producing an enhanced concept over the heart disease prediction within novel data mining techniques; SVM, RF, NB, MLP and j48 the weighted association classifier.

V. APPROACH METHODOLOGY

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

5.1 Classification Algorithms

Classification is a supervised learning procedure that is used for predicting the outcome from existing data. This paper proposes an approach for the diagnosis of heart disease using

Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes 0 for failure and 1 for success[3].

NaÃ¯ve Bayes

NaÃ¯ve Bayes classifier is a supervised algorithm. It is a simple classification technique using Bayes theorem. It assumes independence among attributes. Bayes theorem is a mathematical concept that is used to obtain the probability. The predictors are neither related to each other nor have correlation to one another[7]. All the attributes independently contribute to the probability to maximize it. Many complex real-world situations use Naive Bayes classifiers P(X/Y) = P(Y/X)Ã—P(X)/P(Y),

P(X/Y) is the posterior probability, P(X) is the class prior probability, P(Y) is the predictor prior probability, P(Y/X) is the likelihood, probability of predictor.

P(X/Y) is the posterior probability, P(X) is the class prior probability, P(Y) is the predictor prior probability, P(Y/X) is the likelihood, probability of predictor.

Support Vector Machine

An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized[8]. The goal of SVM is to divide the datasets into classes to find a maximum marginal.

K-Nearest Neighbour

The K-Nearest Neighbour algorithm is a supervised classification algorithm method. It classifies objects dependant on nearest neighbour. It is a type of instance based learning. The calculation of distance of an attribute from its neighbours is measured using Euclidean distance. It uses a group of named points and uses them on how to mark another point. The data are clustered based on similarity amongst them[6]. K-NN algorithm is simple to carry out without creating a model or making other assumptions. This algorithm is versatile and is used for classification, regression, and search. Even though K-NN is the simplest algorithm, noisy and irrelevant features affect its accuracy[13].

Random Forest

Random Forest algorithm is a supervised classification algorithmic technique. In this algorithm, several trees create a forest. Each individual tree in random forest lets out a class expectation and the class with most votes turns into a model's forecast[9]. In the random forest classifier, the more the number of trees higher is the accuracy. It is used for classification as well as regression task, but can do well with classification task, and can overcome missing values. Besides, being slow to obtain predictions as it requires large data sets [16].

XGBoost

XGBoost is an opimized distributed gradient model designed to be highly efficient, flexible and portable. It is a decisiontree based ensemble Machine Learning algorithm that uses gradient boosting framework. It provides an optimized gradient boosting algorithm through parallel processing, tree pruning, handling missing values and regularization to avoid overfitting or bias[7].

Gradient Boosting

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

Gradient boosting trees can be more accurate than random forests. Because we train them to correct each other's errors, they're capable of capturing complex patterns in the data.

Algorithm Used: Gradient Boosting classifier

Boosting is an ensemble method that combines several weak learners into a strong learner sequentially. In boosting methods, we train the predictors sequentially, each trying to correct its predecessor.

Gradient Boosting

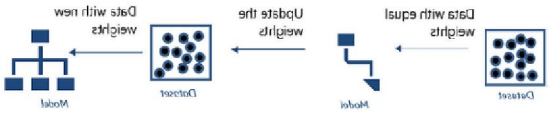
Gradient Boosting is the grouping of Gradient descent and Boosting. In gradient boosting, each new model minimizes the loss function from its predecessor using the Gradient Descent Method. This procedure continues until a more optimal estimate of the target variable has been achieved

Unlike other ensemble techniques, the idea in gradient boosting is that they build a series of trees where every other tree tries to correct the mistakes of its predecessor tree.

Components of Gradient Boosting

- Loss function
- Weak Learners
- Additive Component

STEPS TO GRADIENT BOOSTING CLASSIFICATION



Gradient Boosting Model

STEP 1: Fit a simple linear regression or a decision tree on data [$\boldsymbol{x} = \boldsymbol{input}$, $\boldsymbol{y} = \boldsymbol{output}$]

STEP 2 : Calculate error residuals by subtracting predicted target value from actual target value. [$\boldsymbol{e1} = \boldsymbol{ytrue} - \boldsymbol{ypredicted1}$]

STEP 3 : Fit a new model on the error residuals as the target variables keeping the input variables same.[$\boldsymbol{epredicted1}$]

STEP 4 : Add the predicted residuals to previous predictions [$\boldsymbol{ypredicted2} = \boldsymbol{ypredicted1} + \boldsymbol{epredicted1}$]

STEP 5 : Fit the next model on the remaining residuals. [$\boldsymbol{e2} = \boldsymbol{ytrue} - \boldsymbol{ypredicted2}$]

Repeat steps 2 to 5 until the model starts overfitting or there is no change in the residuals sum

Example

Step 1: **Make initial guess using log of the odds of target variable.**

$$odds = \log\left(\frac{P(Y=1)}{P(Y=0)}\right) = \log\left(\frac{3}{1}\right) = \log(3)$$

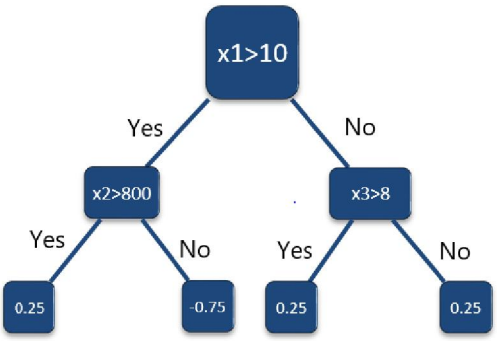
To do classification, we apply softmax transformation.

$$P(Y=1) = \frac{e^{odds}}{1+e^{odds}} = \frac{3}{1+3} = 0.75$$

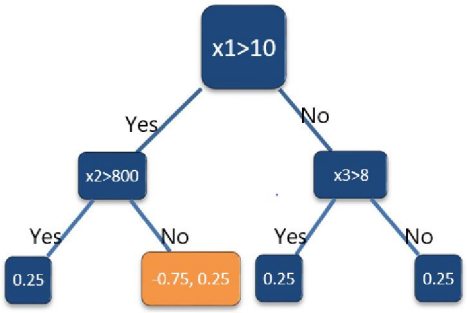
Step 2: **Calculate error residuals or pseudo residuals by subtracting prediction from the observed values**

Step 3: **Compute Classification tree.**

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023



This is an example of classification tree with just two leaves. However, Gradient Boosting often has more than 5 leaves and many leaves can have multiple values. Therefore, Gradient Boosting uses transformation for Classification. Consider the following tree:



Therefore, the value of second leaf is given by the following transformation

$$= \frac{\sum Residual}{\sum[Previous\ Probability * (1 - Previous\ Probablity)]}$$

$$= \frac{-0.75 + 0.25}{0.75(1 - 0.75) + 0.75(1 - 0.75)} = -1.33$$

Step 4: Make the prediction.

$$y_{prediction} = odds + learning_{rate} * residual$$

$$y_{prediction} = \log(3) + 0.1 * (-1.33) = 0.965$$

Learning rate defines the contribution of the new tree. Now new log odds prediction can be converted to probability using softmax function.

$$P(Y = 1) = \frac{e^{0.965}}{1 + e^{0.965}} = 0.724$$

As you can see, the probability has diminished from previous log-odds ratio.

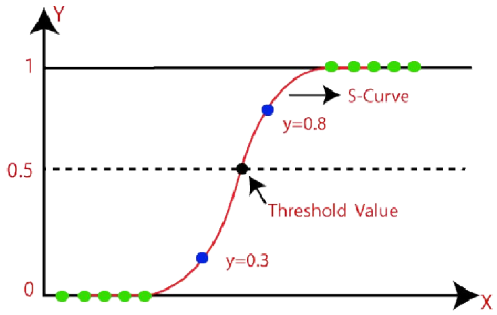
Step 5: Repeat steps until the model starts over-fitting or there is no change in the residuals sum.

Logistic Regression

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).



Logistic Function (Sigmoid Function):

The sigmoid function is a mathematical function used to map the predicted values to probabilities.

It maps any real value into another value within a range of 0 and 1.

The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.

In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Logistic Regression Equation:

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

We know the equation of the straight line can be written as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}$$; 0 for y= 0, and infinity for y=1

Type of Logistic Regression used in this model Binomial**:** In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.

VI. DATA COLLECTION

A data set is an assortment of data. In tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set. Here we have collected data from Kaggle.com. The dataset collected has 14 columns and 421 rows corresponding to 14 medical attributes of 1400 patients.

Sl. No.	Attributes	Range	Type	Description
1.	Age		Numerical	Age in years
2.	Sex	1=Male, 0=Female	Numerical	Describes the gender.
3.	ChestPain	0=Typical angina 1=atypical angina 2=non-anginal pain 3=symptomatic	Numerical	The information about chest pain of a person.
4.	RestBP		Numerical	Resting blood pressure in mm Hg

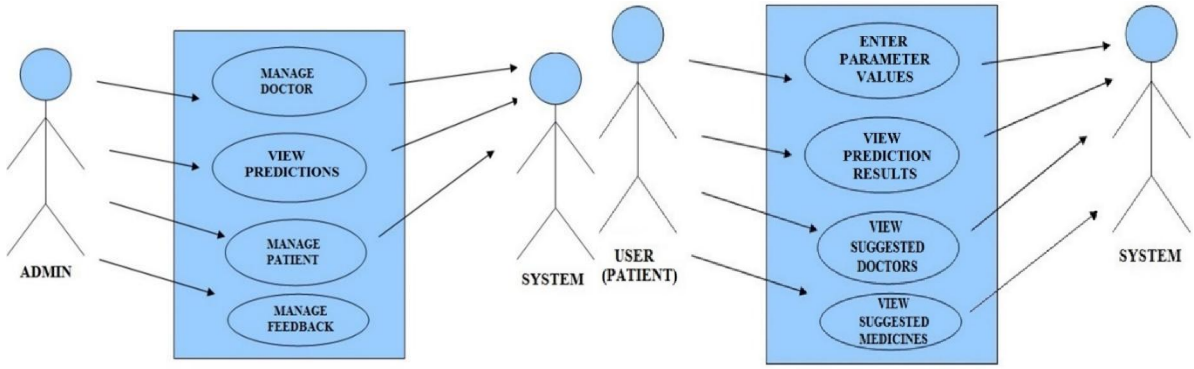
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

5.	Chol		Numerical	Serum cholesterol in mg/dl
6.	Fbs	1 = True 0 = False	Numerical	Fasting blood sugar > 120 mg/dl
7.	RestECG	0=Normal 1=having ST-T wave abnormality 2=Probable or definite left ventricular hypertrophy	Numerical	Resting electrocardiographic results
8.	MaxHR		Numerical	Describes max heart rate
9.	ExAng	1 = Yes 0 = No	Numerical	Describe the information about exercise induced angina
10.	Oldpeak		Numerical	ST depression induced by exercise relative to rest.
11.	Slope	0=Up Sloping 1=Flat 2=Down Sloping	Numerical	The slope of the peak exercise ST segment
12.	Ca		Numerical	Number of major vessels (0-3) coloured by fluoroscopy.
13.	Thal	3 = Normal = Fixed defect = Reversable defect	Numerical	Thalassemia /Blood disorder levels
14.	AHD	= No = Yes	Numerical	Heart disease

Table 1: Attributes Description

Use Case Diagram:

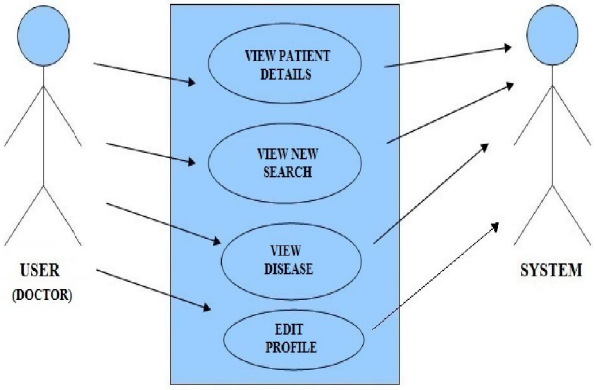
Use case diagram consists of use cases and actors and shows the interaction between them. The key points are:
The main purpose is to show the interaction between the use cases and the actor.
To represent the system requirement from user's perspective.
The use cases are the functions that are to be performed in the module.



Use Case Diagram between ADMIN and SYSTEM:

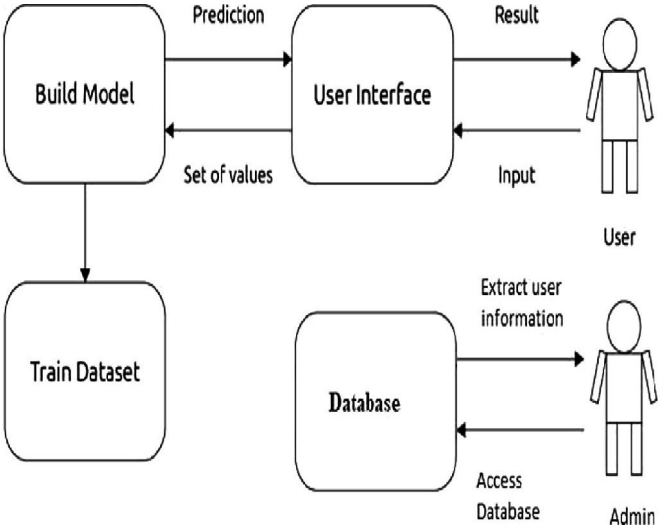
Use Case Diagram between user and system

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

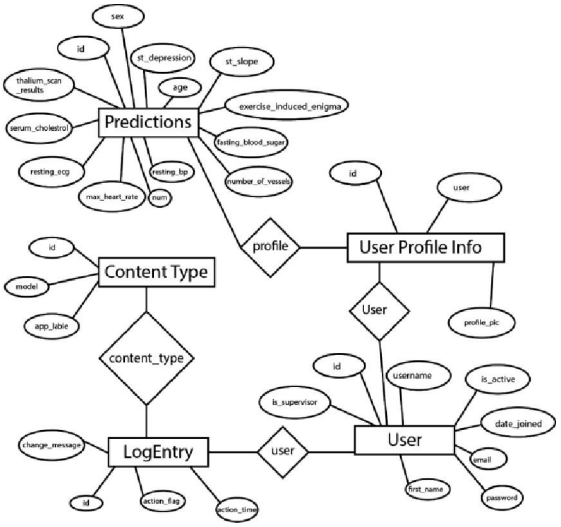


Use Case Diagram between user and system

DFD (Data Flow Diagram)

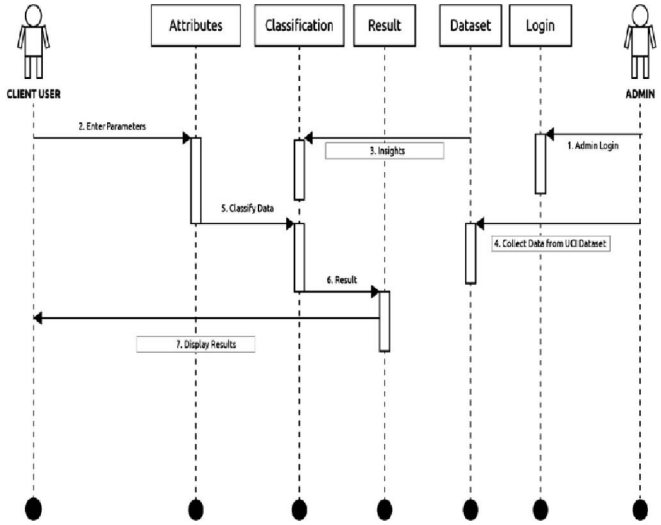


Entity Relationship Diagrams (ER-Diagrams):

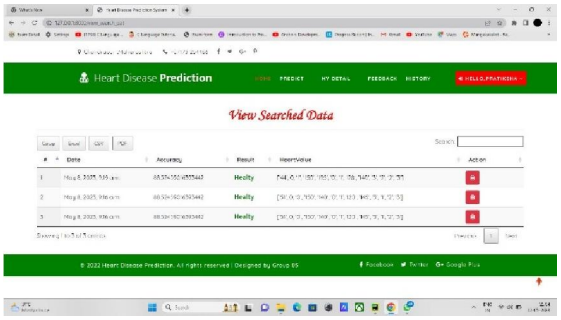


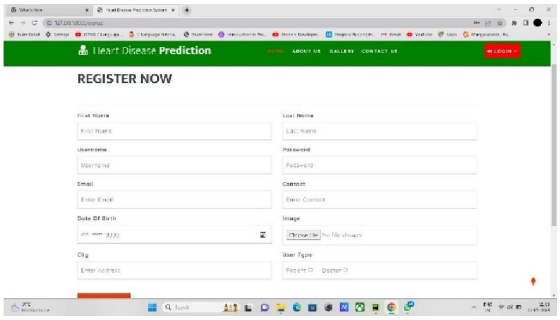
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

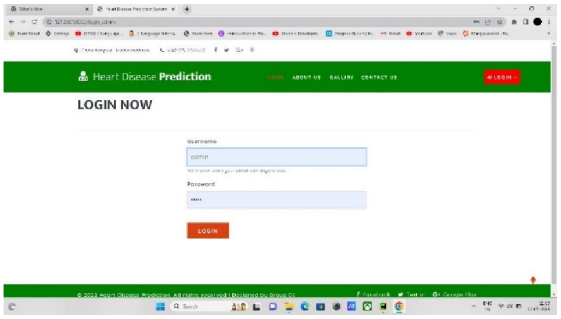
Activity Diagram:

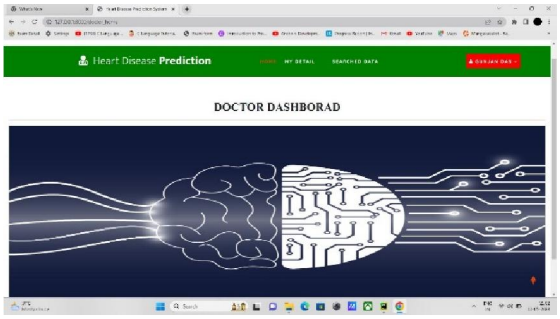


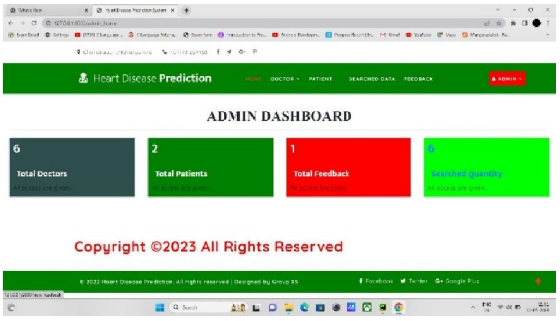
OUTPUT

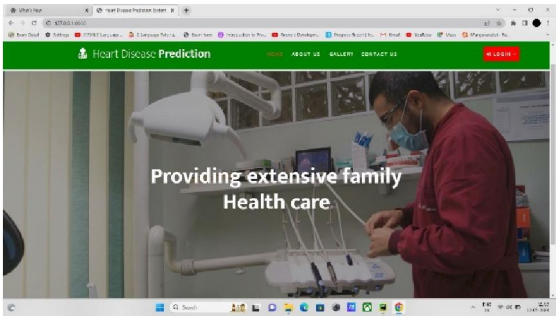




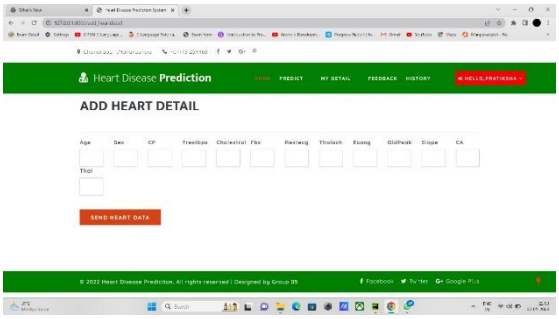








IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023



VII. RESULT AND ANALYSIS

The aim of this research is to analyse the performance of various classification algorithms and in doing so find the most accurate algorithm for predicting whether a patient would develop and heart disease or not. This research was done using techniques of Logistic Regression, NaÃ¯ve Bayes, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Random Forest, XGBoost on the UCI dataset[4]. Dataset was split into training and test data and models were trained and the accuracy was noted using Python

VIII. CONCLUSION

The overall aim is to define various data mining techniques useful in effective heart disease prediction. Efficient and accurate prediction with a lesser number of attributes and tests is the goal of this research. The data were pre-processed and then used in the model[16]. Random Forest with 86.89% and XGBoost with 78.69% are the most efficient algorithms. However, KNearestNeighbor performed with the worst accuracy with 57.83%. We can further expand this research incorporating other data mining techniques such as time series, clustering and association rules and other ensemble techniques. Considering the limitations of this study, there is a need to implement more complex and combination of models to get higher accuracy for early prediction of heart disease[13].

REFERENCES

[1]. Seckeler MD, Hoke TR. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. Clin Epidemiol. 2011;3:67.
[2]. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine- learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE. 2017;12(4):e0174944.
[3]. Ramalingam VV, Dandapath A, Raja MK. Heart disease prediction using machine learning techniques: a survey. Int J Eng Technol. 2018;7(2.8):6847
[4]. 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.
[5]. 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: 22783075, Volume-8 Issue-3, January 2019.
[6]. Internet source [Online].Available (Accessed on May 1 2020): http://acadpubl.eu/ap
[7]. H. Jindal, S. Agrawal, R. Khera, R. Jain and P. Nagrath, “Heart disease prediction using machine learning algorithms”, ICCRDA 2020, IOP Conf. Series: Materials Science and Engineering, 1022 (2021) 012072, DOI:10.1088/1757-899X/1022/1/012072.
[8]. P. Motarwar, A. Duraphe, G. Suganya, M. Premalatha, “Cognitive Approach for Heart Disease Prediction using Machine Learning”, International Conference on Emerging Trends in
[9]. Information Technology and Engineering (ic-ETITE), IEEE, 2020, DOI: 10.1109/icETITE47903.2020.242.
[10]. V. Sharma, S. Yadav, M. Gupta, “Heart Disease Prediction using Machine Learning
[11]. Techniques”, 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), IEEE, 18-19 Dec. 2020, DOI: 10.1109/ICACCCN51052.2020.9362842.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

[12]. A. Nikam, S. Bhandari, A. Mhaske, S. Mantri, "Cardiovascular Disease Prediction Using Machine Learning Models" IEEE Pune Section International Conference (PuneCon), IEEE, 16-18 Dec. 2020, DOI: 10.1109/PuneCon50868.2020.9362367.
[13]. S. Mohan, C. Thirumalai, G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", IEEE Access, Special Section on Smart Caching, Communications, Computing and Cybersecurity for Information-Centric Internet of Things, 2019, DOI: 10.1109/ACCESS.2019.2923707.
[14]. D. Kumar, S. Kumar, K. Arumugaraj, V. Mareeswari, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms", IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, 2018.
[15]. A. Gavhane, G. Kokkula, I. Pandya, K. Devadkar, "Prediction of Heart Disease Using Machine Learning", Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology, IEEE Conference, 2018.
[16]. Minas A. Karaolis,Joseph A. Moutiris, Demetra Hadjipanayi, Constantinos S. Pattichis, Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees, IEEE Transactions on Information Technology in Biomedicine, Vol. 14, N2010
[17]. Sonali. B. Maind, Priyanka Wankar," Research Paper on Basic of Artificial Neural Network", International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), Vol. 2, No. 1,
January 2014, pp. 96-100.
[18].Harleen Kaur , Siri Krishan Wasan and Vasudha Bhatnagar, "The Impact of Data Mining Techniques on Medical Diagnostics ", Data Science Journal, Vol. 5, October 2006, pp. 119126.
[19].R. Dybowski and V. Gant, "Clinical Applications of Artificial Neural Networks", Cambridge University Press, 2007