IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 11, November 2021

DOI: 10.17148/IARJSET.2021.81133

Prediction of Cardiovascular Disease Using Machine Learning Techniques

Chetana Patil[1], Dr .Dinesh. D .Patil[2], Dr. Priti Subramanium[3]

[1]M. Tech Student, Computer Science, SSGBCOET, Bhusawal, India

[2]Head & Associate Professor, Computer Science, SSGBCOET, Bhusawal, India

[3]Assistant Professor, Computer Science, SSGBCOET, Bhusawal, India

Abstract: The medical and health-care industries gather vast volumes of data that could include some secret knowledge that can help make better decisions. Advanced data mining methods are used to provide appropriate results and make successful data decisions. A neural network is used in an efficient cardiovascular disease prediction method to predict the risk level of heart disease. For prediction, the device uses a variety of medical parameters such as age, sex, blood pressure, cholesterol, and obesity. The cardiovascular disease prediction method forecasts the risk of heart disease in patients. Relationships between people are an example of significant understanding. It allows medical causes and trends linked to heart disease to be identified. As a training algorithm, we used a multilayer perceptron neural network with back propagation. The obtained results show that the designed diagnostic device is capable of accurately predicting the risk of heart disease.

Keywords: Genetic Algorithm, Data mining, Naive Bayes, Multilayer perceptron neural network, Machine Learning, Deep Learning, Neural network.

I. INTRODUCTION

Many contributing risk factors, such as high cholesterol, irregular pulse rhythm, diabetes, high blood pressure, and many others, make it difficult to detect cardiovascular disease. Various data mining, machine learning and neural network methods have been used to determine the magnitude of cardiovascular disease in humans. Nave Bayes (NB) and the Genetic Algorithm (GA). It must be treated with caution since the importance of cardiovascular disease is nuanced. Failure is resulting into premature death or can harm the heart. Health research, machine learning and data mining was used to discover various types of metabolic syndromes. Decision trees have also been used to calculate the accuracy of incidents linked to cardiovascular disease [5]. For the prediction of cardiovascular disease, numerous methods of information abstraction have been used, contains data mining methods which are medically proven. Several researches were conducted in order to build a prediction model using not only various techniques, but also by encapsulating two or more techniques. For classification, a dataset with a radial basis function network (RBFN) is used, with more than 70% of the detail used for training data and the remaining more than 30% for classification. In medical science, we also developed the Computer Aided Decision Support System (CADSS). Previous research has shown that using machine learning and data mining techniques in the healthcare industry reduces the time it takes to detect disease and produces more reliable outcomes. We consider using the Genetic Algorithm (GA) to diagnose cardiovascular disease [2]. This method uses effective association rules inferred with the Genetic Algorithm (GA) for selection of tournament, the mutation which gives results in the new proposed fitness method for experimental validation and crossover. We use the well-known dataset which is collected from a UCI machine learning repository we'll see how our research stack up against some of the more well-known knowledge supervised learning methods later on [8]. The most powerful evolutionary algorithm Particle Swarm Optimization (PSO) is introduced and some important rules are implemented for cardiovascular disease. Age, pulse rate, sex, and many others. Heart disease (CVD) generally refers to condition that includes blocked blood vessels or narrowed that can lead to a Myocardial infarctions (heart attack), stroke or chest pain (angina). The term heart disease is sometimes called interchangeably with the term cardiovascular disease [9].

II. LITERATURE SURVEY

The collected and recognized data can be used by social insurance directors to demonstrate indicators of progress administrations. In countries such as India and the United States, coronary disease was the leading cause of death. Naive Bayes experiences different forms of heart-related issues when performing machine learning calculations such as a Logistic relapse, irregular woods, angle boosting, and Support vector machine and order calculations, for example. These measurements can be used to improve information storage for practical and legal purposes [4].

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 8, Issue 11, November 2021
DOI: 10.17148/IARJSET.2021.81133

2.1 Logistic Regression

For binary classification, logistic regression is well-known, and it is one of the most effective machine learning algorithms. Because of its simplicity, it can be applied to a wide variety of problems and offers appropriate solutions [7]. It operates on a categorical dependent variable. Binary dependent variables such as 0s or 1s, pass or fail are used [6].

2.2 Naive Bayes

The Navies Bayes classifier, which employs the Bayesian algorithm, is one of the best classification algorithms in machine learning. The Naive Bayes classification algorithm is highly scalable, and it necessitates linear variables in the form of predictor variables in the problem statement. It's the same for classification and regression, and it's difficult [2].

2.3. Support Vector Machine

Another classification technique is Support Vector Machines which separates data values by the creation of hyper planes. Hyper planes can be of various shapes based on the spread of data, but only those points which help in differentiating between the classes are considered for classification [3]**.**

2.4. Kernel Functions

If data points are in nonlinear fashion, the kernel function makes them in linear decision plane. Some Kernel functions are as follows:

a. Linear Function: In these type of kernel the hyper surface is a straight line. Best results are provided by Linear Kernel functions for classifiers which are exactly two target classes.

b. Radial Basis Function: When points cannot be separated in a linear fashion that time it uses Radial Basis Function. The function bring points into a various shape generally circular/radial fashion to perform further actions.

c. Polynomial Function: In this type of kernel functions the hyper plane is mostly a polynomial like hyperbola, parabola [8].

2.5 Random Forest

Random forest is a regression and classification machine learning algorithm. For each parameter, it implements a decision tree [1]. It corrects their training data sets over fitting. By following the steps of data pre-processing and data interpretation, it also prevents outliers and missing values. It is a machine learning approach that encapsulates weak models to develop a dynamic model. The random forest tree depicts the different decision trees associated with the model [5]. The American cardiovascular Association's Statistics Committee always tracks and reviews sources of data on cardiovascular disease and stroke in the country so that the annual Statistical Update contains the most up-to-date information. The United states 2021 Statistical Update is the product of a year's worth of work by scientists, volunteer physicians, American Heart Association employees and government professionals [10]. This year's version provides informative data on population-level heart health monitoring and benefits, as well as a greater research on adverse pregnancy outcomes, social determinants of health, the global burden of cardiovascular disease, vascular contributions to brain health and more evidence-based approaches to improving cardiovascular disease-related behaviours[9]. Meta-regression and Subgroup analyses are two strategies that can be used to predict the investigation of treatment effects associated with the discrepancies between studies and causes of heterogeneity [10]. Subgroup analyses evaluate and divide data according to variables of interest that were predicted a priori during the initial protocol stage. However, the estimates and their confidence intervals for all subgroups should be used to interpret the analyses [8]. In the medical_elds directly related to this article, there is lots of related work. In the ANN, medical _eld has been developed to achieve the highest accuracy prediction [8]. Cardiovascular disease is forecasted using ANN's back propagation multilayer perception (MLP). When the obtained results are compared to the results of existing models in the same domain, they are found to be superior [6]. Patterns are discovered using DT, NN, Support Vector Machines (SVM), and Naive Bayes on data from cardiovascular disease patients obtained at the UCI laboratory. With these algorithms, the efficiency and accuracy of the results are compared. The proposed hybrid function competes with other current methods [9] by returning results of 86:8 precent for F-measure. Without segmentation the Convolutional Neural Networks (CNN) classification is added. During the training data process, the Electrocardiogram (ECG) signals are used to identify and calculate heart cycles with different initial locations. In the patient's testing data process, CNN can produce features in a variety of positions [5], [6]. Previously, a significant amount of data produced by the medical industry was not effectively utilized. The new functions presented here are simple and successful in lowering the cost and improving the forecasting of cardiovascular disease prediction and classification using data mining and machine learning (ML) and deep learning (DL) techniques are highly accurate [4].

III. PROPOSED SYSTEM

The patient information is included in the datasets. The attributes that are useful for the prediction of heart disease are chosen by attribute collection. Following the identification of data from available resources, it is further selected for

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 8, Issue 11, November 2021
DOI: 10.17148/IARJSET.2021.81133

processing, which involves data cleaning and noise reduction (i.e. missing data). To determine the likelihood of developing heart disease, different classification algorithms are applied to the pre-processed data. It also determines algorithm accuracy and compares algorithm accuracy across all algorithms [15].

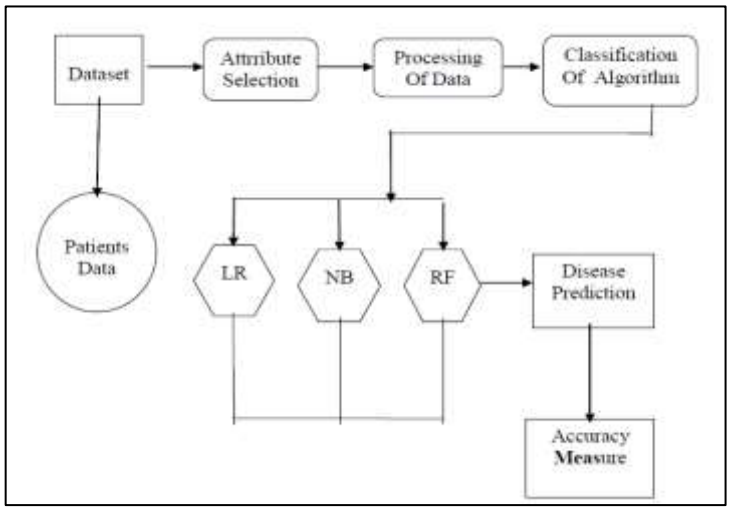

Fig. 1 Proposed System

IV. HIGH LEVEL DESIGN

4.1 High Level Design

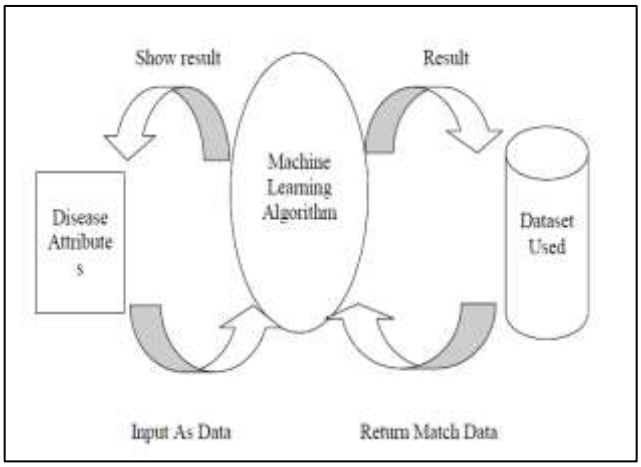

Fig. 2 High Level Design

V. DATA MINING TECHNIQUES RESULTS

All input parameters in terms of their significance based on the results obtained from each technique using various functions as.

TABLE I DATA MINING TECHNIQUES RESULTS

Accuracy	Specificity	Sensitivity	Method
0.8433	0.895	0.7826	SVM
0.81	0.8395	0.7753	KNN K=7
0.8366	0.8641	0.8043	Naive Bayes
0.79	0.8148	0.7608	Decision Tree

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 11, November 2021

DOI: 10.17148/IARJSET.2021.81133

VI. DATA PRE-PROCESSING

After collection of various records cardiovascular disease data is pre-processed. The dataset includes a total of some patient records, where some records are with few missing values. Those records have been removed from the dataset and the remaining patient records are used in pre-processing. The binary classification and multiclass variable are introduced for the parameters of the given dataset. The multi-class variable is used to identify the presence or absence of cardiovascular disease. The pre-processing of data is used out by converting medical records into diagnosis values [14]. The results of pre-processing data is for remaining patient records indicate that some records show the value of 1 establishing the presence of cardiovascular disease while the remaining reflected the value of 0 indicating the absence of cardiovascular disease [11].

TABLE II UCI DATASET ATTRIBUTES DETAILED INFORMATION

Attribute	Description	Type
Age	Patient's age in completed years	Numeric
Sex	Patient's Gender (male represented as1 and female as 0)	Nominal
Cp	The type of Chest pain categorized into 4 values: 1. typical angina, 2. atypical angina, 3. non-anginal pain and 4. asymptomatic	Nominal
Trestbps	Level of blood pressure at resting mode (in mm/Hg at the time of admitting in the hospital)	Numeric
Chol	Serum cholesterol in mg/dl	Numeric
FBS	Blood sugar levels on fasting > 120 mg/dl; represented as 1 in case of true, and 0 in case of false	Nominal
Resting	Results of electrocardiogram while at rest are represented in 3 distinct values: Normal state is represented as Value 0, Abnormality in ST-T wave as Value 1, (which may include inversions of T-wave and/or depression or elevation of ST of > 0.05 mV) and any probability or certainty of LV hypertrophy by Estes' criteria as Value 2	Nominal
Thali	The accomplishment of the maximum rate of heart	Numeric
Exang	Angina induced by exercise. (0 depicting 'no' and 1 depicting 'yes')	Nominal
Oldpeak	Exercise-induced ST depression in comparison with the state of rest	Numeric
Slope	ST segment measured in terms of the slope during peak exercise depicted in three values: 1. unsloping, 2. flat and 3. downsloping	Nominal
Ca	Fluoroscopy coloured major vessels numbered from 0 to 3	Numeric
Thal	Status of the heart illustrated through three distinctly numbered values. Normal numbered as 3, fixed defect as 6 and reversible defect as 7.	Nominal
Num	Heart disease diagnosis represented in 5 values, with 0 indicating total absence and 1 to 4 representing the presence in different degrees.	Nominal

TABLE III UCI DATASET RANGE AND DATA TYPE

AGE	Numeric [29 to 77;unique=41;mean=54.4;median=56]
SEX	Numeric [0 to 1;unique=2;mean=0.68;median=1]
CP	Numeric [1 to 4;unique=4;mean=3.16;median=3]
TESTBPS	Numeric [94 to 200;unique=50;mean=131.69;median=130]
CHOL	Numeric [126 to 564;unique=152;mean=246.69;median=241]
FBS	Numeric [0 to 1;unique=2;mean=0.15;median=0]
RESTECG	Numeric [0 to 2;unique=3;mean=0.99;median=1]
THALACH	Numeric [71 to 202;unique=91;mean=149.61;median=153]
EXANG	Numeric [0 to 1;unique=2;mean=0.33;median=0.00]
OLPEAK	Numeric [0 to 6.20;unique=40;mean=1.04;median=0.80]
SLOPE	Numeric [1 to 3;unique=3;mean=1.60;median=2]
CA	Categorical [5 levels]
THAL	Categorical [4 levels]
TARGET	Numeric [0.00 to 4.00;unique=5;mean=0.94;median=0.00]

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 11, November 2021

DOI: 10.17148/IARJSET.2021.81133

VII. COMPARATIVE RESULTS OF SPECIFICITY, SENSITIVITY, ACCURACY

This section compares the specificity, sensitivity, accuracy, and some of the employed techniques in terms of their confusion matrices. Support Vector Machine, Naïve Bayes, Decision tree, and K-Nearest Neighbour were applied to the dataset. Table 2 shows the sensitivity, specificity and accuracy of these data mining and machine learning techniques [13].

79% to 84.33% are the Accuracy ranges for cardiovascular disease prediction. SVM achieved the highest accuracy (84.33%) according to Table 1. SVM and Naïve Bayes achieved the highest accuracy, then KNN (k=7 resulted in the best accuracy as compared to other predicted values) and decision tree, respectively. IBM SPSS Modellers and Weka used for developing the data mining techniques [12].

TABLE IV CARDIOVASCULAR DISEASE DATASET INDICES

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	1=male
		0=female
Cp	Discrete	Chest pain type:
		1-typical angina
		2-atypical angina
		3-non-angina pain
		4-asymptomatic
Fbs	Discrete	FBS>120 (mg/dl)
		1=Yes
		0=No
Restecg	Discrete	Resting electrocardiographic results:
		0=Normal
		1-having ST-T wave abnormality
		2=showing probable or defined left ventricular hypertrophy
Trestbps	Continuous	Resting blood pressure in (mm Hg)
Exang	Discrete	Exercise-induced angenia:
		1=Yes
		0=No
Chol	Continuous	Serum cholesterol in (mg/dl)
Ca	Discrete	Number of major vessels coloured by fluoroscopy that range between 0 and 3
Diagnosis	Discrete	Diagnosis classes:
		1=healthy
		2=patient who is subject to possible heart disease

VIII. CONCLUSION

Cardiovascular disease prediction function has been presented by using data mining and machine learning techniques. The framework is implemented using a neural network, Genetic and Naive Bayes and MLP algorithms. The MLP Model implements better results and assists medical professionals and even domain experts in planning for a earlier and better prediction and diagnosis for patients. This research's future course can be carried out using different machine learning and data mining techniques to improve prediction techniques. Furthermore, new feature selection methods can be introduced to obtain a detailed understanding of the important features and improve cardiovascular disease prediction results. Instead of using simulations and theoretical approaches, real-world datasets are used. The proposed hybrid HRFLM solution encapsulates Random Forest and HRFLM characteristics.

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 11, November 2021

DOI: 10.17148/IARJSET.2021.81133

REFERENCES

[1] Rosamond W, Flegal K, Furie K, et al. Heart disease and stroke statistics 2008 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Circulation. 2008; 117(4):e25–146. [PubMed].

[2] National Heart Lung and Blood Institute Fact Book, Fiscal Year 2006. Bethesda, Md: National Heart Lung and Blood Institute, National Institutes of Health; 2006. [12 April 2011]. Last accessed at http://www .nhlbi.nih.gov/about/factbook-06/toc.htm on.

[3] Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). JAMA. 2001; 285(19):2486–97. [PubMed].

[4] Hayden M, Pignone M, Phillips C, et al. Aspirin for the primary prevention of cardiovascular events: a summary of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med. 2002; 136 (2): 161–72. [PubMed].

[5] Sharma Purushottam, Dr. Kanak Saxena, Richa Sharma, "Heart Disease Prediction System Evaluation using C4.5 Rules and Partial Tree", Springer, Computational Intelligence in Data Mining,vol.2, 2015, pp.285-294.

[6] S.Prabhavathi, D.M.Chitra, "Analysis and Prediction of Various Heart Diseases using DNFS Techniques", International Journal of Innovations in Scientific and Engineering Research, vol.2, 1, January 2016, pp.1-7

[7] Tina R. Patil, Mrs. S.S. Sherekar, " Performance Analysis of Naïve Bayes and J48 Classification algorithm for Data Classification" , International Journal Of Computer Science and Applications, Vol. 6, No.2, Apr 2013..

[8] S.Kumar Mandal, Animesh Hazra, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee, "Heart Diseases Diagnosis and Prediction Using Machine Learning and Data Mining Techniques:A Review", Advances in Computational Sciences and Technology, Vol. 10, No.7, July-2017.

[9] Heart Disease Diagnosis Using Data Mining Techniques Ramin Assari1 , Parham Azimi2 and Mohammad Reza Taghva1 1Department of IT Management, Allameh Tabataba'i University, Tehran, Iran 2Faculty of Mechanical Engineering and Industrial Engineering, Islamic Azad University, Qazvin, Iran

[10]Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques senthilkumar mohan 1 , chandrasegar thirumalai1 , and gautam srivastava 2,3, (Member, IEEE) 1School of Information Technology and Engineering, VIT University, Vellore 632015, India 2Department of Mathematics and Computer Science, Brandon University, Brandon, MB R7A 6A9, Canada 3Research Center for Interneural Computing, China Medical University, Taichung 40402, Taiwan.

[11]J. Wu, M. Dong, K. Ota, J. Li, and Z. Guan, ''FCSS: Fog computing based content-aware filtering for security services in information centric social networks,'' IEEE Trans. Emerg. Topics Comput., to be published. doi: 10.1109/TETC.2017.2747158.

[12] Y. Meidan, M. Bohadana, A. Shabtai, J. D. Guarnizo, M. Ochoa, N. O. Tippenhauer, and Y. Elovici, ''ProfilIoT: A machine learning approach for IoT device identification based on network traffic analysis,'' in Proc. Symp. Appl. Comput., Apr. 2017, pp. 506–509.

[13]Guizhou Hu, Martin M. Root, "Building Prediction Models for Coronary Heart Disease by Synthesizing Multiple Longitudinal Research Findings", European Science of Cardiology, 10 May 2005.

[14]Marjia Sultana, Afrin Haider, "Heart Disease Prediction using WEKA tool and 10-Fold cross-validation", The Institute of Electrical and Electronics Engineers, March 2017

[15][6] Mr. Chala Beyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", International Journal of Pure and Applied Mathematics, 2018.

[16]SENTHILKUMAR MOHAN 1, CHANDRASEGAR THIRUMALAI1, AND GAUTAM SRIVASTAVA 2, 3, (Member, IEEE), "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" Research Center for Interneural Computing, China Medical University, Taichung 40402, Taiwan.

