

PAPER • OPEN ACCESS

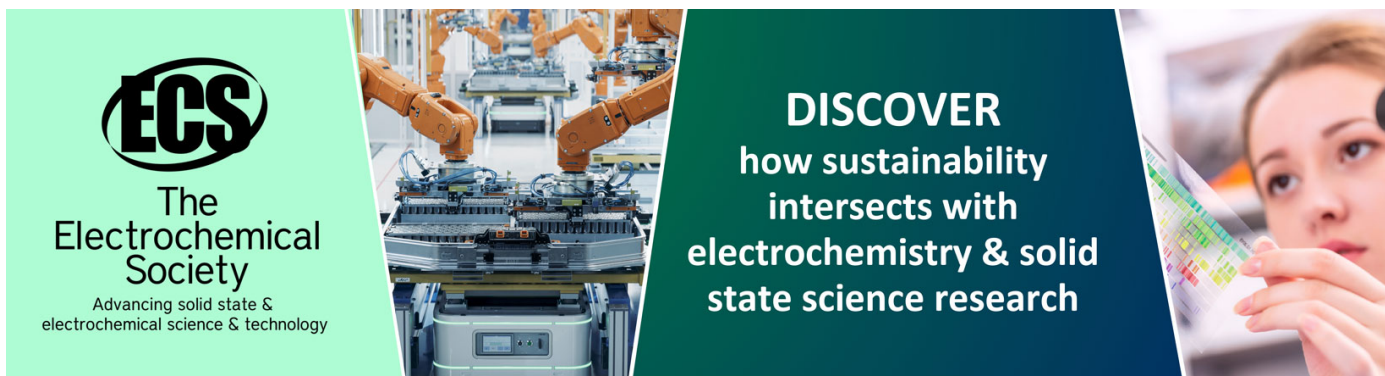
## HEART DISEASE CLASSIFICATION USING MACHINE LEARNING TECHNIQUES

To cite this article: R. Radhika and S. Thomas George 2021 *J. Phys.: Conf. Ser.* **1937** 012047

View the [article online](#) for updates and enhancements.

You may also like

- [Pneumonia identification based on lung texture analysis using modified k-nearest neighbour](#)  
S Kana Saputra, Insan Taufik, Mhd Hidayat et al.
- [An Ensemble Heart Disease Prediction Model Bagged with Logistic Regression, Naïve Bayes and K Nearest Neighbour.](#)  
Annwasha Banerjee Majumder, Somsubhra Gupta and Dharmpal Singh
- [Attribute Weighting Based K-Nearest Neighbor Using Gain Ratio](#)  
A A Nababan, O S Sitompul and Tulus



**ECS**  
The  
Electrochemical  
Society  
Advancing solid state &  
electrochemical science & technology

**DISCOVER**  
how sustainability  
intersects with  
electrochemistry & solid  
state science research

## HEART DISEASE CLASSIFICATION USING MACHINE LEARNING TECHNIQUES

**R. Radhika<sup>1</sup>, S. Thomas George, Associate Professor<sup>2</sup>,**

Department of Biomedical Engineering, Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu, India.

Email: [thomasgeorge@karunya.edu](mailto:thomasgeorge@karunya.edu) , [radhiravi9762@gmail.com](mailto:radhiravi9762@gmail.com)

### ABSTRACT:

Heart disease is one of the Leading reason for death around the world. In which machine learning is a method that predicts the emerging prospects of Heart Disease. Machine learning is used in taking care of numerous issues in information science. The basic utilization of machine learning is the forecast of a result dependent on already existing information. The machine takes the designs from the current dataset, and it is applied on an obscure dataset to foresee the result. Order method in AI is usually used for expectation. Some arrangement calculations foresee with acceptable precision, while others show a restricted exactness. Here, we play out an order dependent on various arrangement calculations like K-Nearest Neighbour, Support Vector Machine, Naïve Bayes, logistic regression, decision tree algorithm and random forest algorithm

**Keywords:** Heart disease, Machine learning, K-Nearest Neighbour, Support Vector Machine, Naïve Bayes, logistic regression, decision tree algorithm, random forest algorithm.

### 1. INTRODUCTION:

An individual can have heart disease and not feel wiped out. A few group with heart disease have indications. This is when there are changes or torment in the body to show a sickness is there. A few manifestations of heart disease are : Pain in the chest, Trouble breathing, Palpitations (an inclination that the heart is pulsating excessively quick), Swelling of feet or legs, Feeling feeble in light of the fact that the body and cerebrum are not getting sufficient blood to supply them with oxygen, Cyanosis (skin turning a blue tone). There are many danger factors for heart illnesses: age, sex, tobacco use, actual idleness, unreasonable liquor utilization, undesirable eating routine, weight, hereditary inclination and family background of cardiovascular sickness, raised pulse (hypertension), raised glucose (diabetes mellitus), raised blood cholesterol (hyperlipidemia), undiscovered celiac infection, psychosocial variables, destitution and low instructive status, and air contamination. While the individual commitment of each hazard factor changes between various networks or ethnic gatherings the general commitment of these danger factors is exceptionally reliable. A portion of these danger factors, like age, sex or family ancestry/hereditary inclination, are unchanging; in any case, numerous significant cardiovascular danger factors are modifiable by way of life change, social change, drug treatment (for instance anticipation of hypertension, hyperlipidemia, and diabetes). Individuals with corpulence are at expanded danger of atherosclerosis of the coronary supply routes. Age is the main danger factor in creating cardiovascular or heart illnesses, with roughly a significantly increasing of hazard with every time of life. Coronary greasy streaks can start to frame in youthfulness. It is assessed that 82% of individuals who pass on of heart disease are 65 and more established. At the same time, the danger of stroke pairs each decade after age 55. Numerous clarifications are proposed to clarify why age expands the danger of cardiovascular/heart infections. One of them identifies with serum cholesterol level. In many populaces, the serum absolute cholesterol level increments as age



increments. In men, this increment levels off around age 45 to 50 years. In ladies, the increment proceeds pointedly until age 60 to 65 years.

Heart disease determination is quite possibly the most basic and testing assignments in the medical services field. It should be analyzed rapidly, productively and accurately to save people's lives. It needs the patient to do numerous tests, and medical care experts should cautiously inspect the outcomes. That is the reason scientists have been keen on anticipating heart disease, and they created diverse heart disease forecast frameworks utilizing different AI calculations. Some of them accomplished preferred outcomes over others. Many used the well-known UCI heart disease dataset to prepare and test their classifier, while others utilized information got from different clinics available to them.

Hence, forestalling Heart illnesses has gotten more than needed. Great information driven frameworks for foreseeing heart illnesses can improve the whole examination and counteraction measure, ensuring that more individuals can carry on with sound lives. AI helps in foreseeing the Heart illnesses, and the expectations made are very precise. It includes investigation of the heart disease patient dataset with appropriate information preparing. At that point, various models were prepared and the expectations are made with various calculations like KNN, SVM, Decision Tree, Naïve bayes, Random Forest, Logistic Regression.

## 2. RELATED WORKS:

Tan et al. depicted a cross breed strategy where two AI calculations like SVM, G.A algorithms were implemented with covering approach. Informational indexes like diabetes infection, bosom malignant growth sickness, heart disease and hepatitis etc. were accommodated from the Irvine UC AI archive for cross- examination. Subsequent to implementing half GA and SVM approach, an exactness of 84.07% is acquired for coronary disease. In most of the diabetes information, 78.26% exactness was accomplished.

Youness Khourdifi et al. portrayed, by contrasting distinctive order models and various calculations better outcomes can be gotten. Each calculation has its inborn ability to beat other calculation relying on the circumstance. For instance, Random Forest executed much superior with an enormous amount of datasets where information is little while Support Vector Machine performs better with fewer informational collections. Finally subsequent to contrasting the various calculations and the proposed streamlined model by ACO, PSO etc., he found K-NN with 99.7 % to be the best as compared to RF with 99.6 %.

Vembandasamy et al. found presence or nonappearance of heart disease by utilizing Naive Bayes classifier. The dataset was acquired from organizations in Chennai and collected records of around 500 patients and had 11 credits (counting the determination) for main diabetic examination. Waikato Environment for Knowledge Analysis apparatus, which is an assortment of ML calculations, is utilized to apply Naive Bayes classifier. 86.4198% accuracy was achieved in their work. Medhekar et al. in [12] implemented a framework which organized the information into five classifications utilizing Naive Bayes classifier. 4-class classifications was done in this method including none, normal, low, high and very high. The methodology predicts any chance of coronary abnormality in the information. Table 1 shows the dataset utilized for testing. The proposed method showed a precision of 88.96%.

Das et al. in implemented a framework utilizing ANN with Ensemble strategy. Dataset shown in the

Table 1 was used for this purpose. The group model gave expanded speculation by joining various models prepared on a similar undertaking. The device used to carry out the test was SAS undertaking excavator 5.2 where the outcomes showed that the model anticipated coronary abnormality with a precision of 89.01%. Chen et al. in [13] built up a heart disease expectation framework (HDPS) utilizing ANN. The Learning Vector Quantization (LVQ) was utilized in this examination. The ANN used in the model has thirteen neurons for the information layer, six neurons for the secret layer and two neurons for the yield layer.

Sabarinathan et al. in utilized the Decision Tree J48 calculation for highlight choice and for anticipating heart disease. Thirteen clinical credits/highlights were used in the database, and 240 clinical records were collected for preparing and rest for testing. The precision accomplished was 75.8333% utilizing every one of the highlights; while the exactness is improved to 76.67% utilizing highlight choice. Moreover, when more superfluous highlights were taken out, the exactness is improved to 85%. The paper asserts that the J48 calculation empowers choosing least highlights to improve expectation precision. Patel et al. in [16] 8 analyzed a few choice tree calculations utilizing WEKA apparatus on the dataset to decide nonappearance and presence of heart disease. The various calculations tried were J48, strategic model-tree, and irregular backwoods. The J48 outperforms all with a precision of 56.76%.

Shouman et al. implemented K-Nearest neighbor (KNN) to foresee coronary abnormality by using the Cleveland dataset. Casting a ballot is the technique for separating the information into modules and applying the classifier to every module. Assessment was finished utilizing 10-overlay cross-approval. The outcomes clearly showed that in absence of casting a ballot, the exactness increased to 97.4% from 94% with different instance for K. At the point when K=7, the precision was the most elevated at 97.4%. Utilizing the democratic procedure, in any case, didn't improve the exactness. The outcomes showed that at K=7, the precision diminished to 92.7%.

Wiharto et al. investigated the precision of SVM calculation on the dataset (UCI) to analyze any coronary abnormality. The investigation included different SVM types like Decision Direct Acyclic Graph (DDAG), Binary Tree Support Vector Machine (BTSVM), One-Against-All (OAA), Exhaustive Output Error Correction Code (ECOC) etc. The dataset acquired was first preprocessed utilizing a scaler (min-max). The following steps includes carrying out the calculation which was finished utilizing the SVM calculations referenced previously. In the exhibition assessment, BTSVM performed better compared to different calculations with 61.86% by and large precision.

Pouriyeh et al. in led a far reaching examination of various characterization strategies on the Cleveland database to figure out the classifier which outperforms the rest. The classifiers included were Single Conjunctive Rule Learner (SCRL), Decision Tree (DT), Multi-layer Perceptron (MLP), Naive Bayes (NB), Radial Basis Function 9 (RBF) etc. The creators utilized the K-Fold Cross Validation procedure to evaluate the best of the classifiers. In each of the individual classifier, the exhibition assessment measurements were exactness, accuracy, review, ROC bend etc. In each of the individual KNN classifier, various estimations of K had a go at, bringing about K=9 as the optimal one. In the ANN, a few neurons were tested to show up at best blend which was thirteen, two and 7. The examination was carried out in two modules: first one included contrasting the various classifiers referenced above, while the subsequent one included applying the gathering methods.

### 3.METHODOLOGY:

#### 3.1. DATA PRE-PROCESSING:

The presentation and exactness of the prescient model isn't just influenced by the calculations utilized, yet in addition with the nature of the dataset and its strategies. This preprocessing stage is vital on the grounds that it readies the dataset and places it in a structure that the calculation gets it. Another important thing is the size in which the dataset varies. Some datasets have numerous ascribes that makes the dataset harder for the calculation to break down it, find examples, or make precise forecasts. Such issues can be addressed by examining the dataset and utilizing the appropriate information preprocessing procedures. Information preprocessing steps incorporates: information cleaning, information change, missing qualities ascription, information standardization, include determination, and different advances relying upon the idea of the dataset.<sup>5</sup> Datasets can have mistaken, missing information, redundancies, clamor, and numerous different issues which cause the information to be unacceptable to be utilized by the AI calculation straightforwardly.

#### 3.2. PERFORMANCE EVALUATION MATRICES:

The measurements referenced beneath are utilized by scientists to assess expectation models and show their presentation results. We give a short definition to every technique without digging into the profound subtleties and numerical conditions.

$$\text{ACCURACY} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

$$\text{PRESICION} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{SENSITIVITY} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{F-MEASURE} = (2 * \text{PRECISION} * \text{RECALL}) / (\text{PRESICION} + \text{RECALL})$$

$$\text{ROC CURVE, } Y = P / (P + N)$$

The most broadly utilized presentation assessment metric is exactness, which is utilized in all examination papers talked about in our article. Subsequently, the focal point of this outline article is on ordering, contrasting and inspecting past work dependent on the exactness.

#### 3.3. HEART DISEASE DATASET:

The dataset that is utilized in most of exploration papers is the heart disease dataset got from the UCI (University of California, Irvine C.A) Center for AI and canny frameworks. It contains four information bases from four clinics. Every data set has similar number of highlights, which is 14, yet various quantities of records. The Cleveland dataset is the most utilized dataset by AI scientists, because of containing less missing ascribes than the other datasets and having more records. The "num" field alludes to the presence of heart disease in the patient. It is number esteemed from 0 (no presence) to 4. The Cleveland dataset contains 303 instances. The different dataset attributes with their values are explained in table 3.3.1.

**Table:3.3.1: DATASET ATTRIBUTES**

ATTRIBUTE	VALUES AND MEANINGS
Age1	Age in year
Gender1	Value 1 and 0 for male and female
Cp1	Pain in chest yes/no
Test bps1	Blood pressure during resting
Chol1	Cholesterol of serum in mg/dl
Fbs1	Blood sugar during fasting
Restecg1	Resting ECG results
Oldpeak1	ST depression induced by exercise relative to test
Slope1	The slope of peak exercise of ST segment  Value 1: unsloping  Value 2: flat  Value 3: down sloping
Ca1	Number of major vessels (0-3) coloured by fluoroscopy
Thal1	3= normal  6= fixed defect
Num1	Diagonal of heart disease  Value 0: no risk  Value 1: low risk  Value 2: risk  Value 3: high risk  Value 4: higher risk

### 3.4. APPLYING VARIOUS MACHINE LEARNING ALGORITHMS

#### 3.4.1. LOGISTIC REGRESSION

This algorithm is used for learning grouping calculation used to foresee the similarity of an objective variable. The main idea of ward variable is dichotomous, which implies either 1 or 0. Numerically, a logistic regression model predicts  $P(Y=1)$  as a component of  $X$ . It is one of the least complex ML calculations that can be utilized for different grouping issues, for example, spam identification, Diabetes expectation, disease location, and so forth

#### 3.4.2. KNN ALGORITHM

This algorithm performs calculation which is straightforward, regulated AI calculation that can be utilized to tackle both arrangement and relapse problem. It is not difficult to execute and see, however has a significant disadvantage of turning out to be essentially eases back as the size of that information being used develops. It utilizes information with a few classes to foresee the characterization of the new example point.

### 3.4.3. SVM

Backing vector machines (SVMs) are incredible yet adaptable directed AI calculations which are utilized both for arrangement and relapse. Be that as it may, by and large, they are utilized in characterization issues. SVMs have their special method of execution when contrasted with other AI calculations.

### 3.4.4. NAÏVE- BAYES

This algorithm performs calculations which mainly depends on Bayes' Theorem. This is nothing but a solitary calculation yet a group of calculation is performed in which every one of them share a typical standard, for example each pair of highlights being arranged is free of one another. Credulous Bayes is a probabilistic AI calculation that can be utilized in a wide assortment of order undertakings. Normal applications incorporate sifting spam, grouping reports, notion forecast and so forth

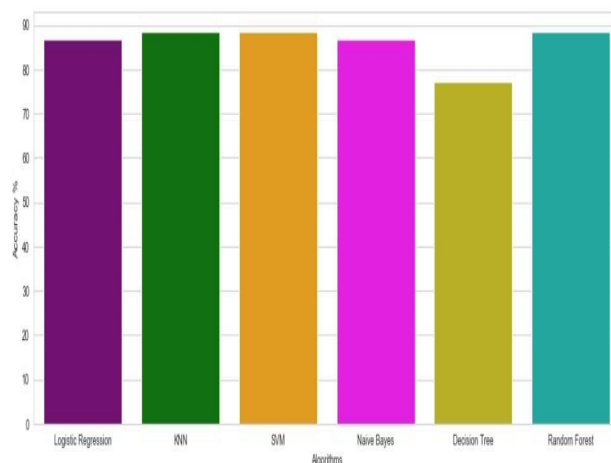
### 3.4.5. DECISION TREE

Decision tree algorithm has a place with the group of regulated learning calculations. In contrast to other regulated learning calculations, the decision tree algorithm can be utilized for tackling relapse and characterization issues as well. The objective of utilizing a Decision Tree is to make a preparation model that can use to foresee the class or estimation of the objective variable by taking in straightforward choice guidelines surmised from earlier data (training information).

### 3.4.6. RANDOM FOREST

Random forest is a group learning strategy for arrangement, relapse and different errands that work by developing a huge number of choice trees at preparing time and the method of the classes (characterization) or mean/normal expectation. It is additionally perhaps the most utilized calculations, due to its straightforwardness and variety.

## MODEL PERFORMANCE COMPARISON



**Fig:3.4:** Model Performance Comparison

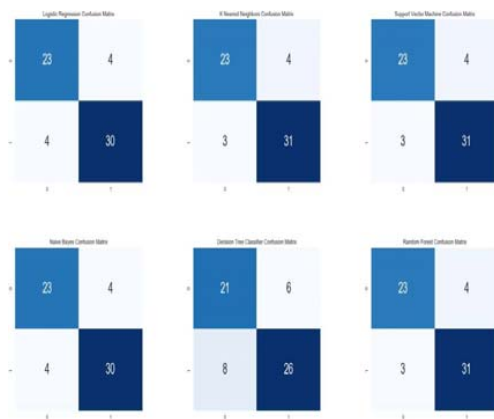
Most of the models work fine but among all the algorithms KNN and Random forest are the best with 88.52 %. Thus the accuracy rate for the following algorithms are explained below in table 3.4.2.

**TABLE:3.4.2: RESULTS AND ACCURACY**

ALGORITHM	ACCURACY (%)
-----------	--------------

LOGISTIC REGRESSION	86.89
KNN ALGORITHM	88.52
SVM ALGORITHM	88.11
NAÏVE BAYES	86.89
DECISION TREE	77.05
RANDOM FOREST	88.52

Then the confusion matrix for the following machine learning algorithms are given below in fig:3.4.2.



**FIG:3.4.2:** Confusion matrix

#### 4. CONCLUSIONS

This task outlines different machine learning grouping strategies for classification of heart disease. Numerous illustrative papers on utilizing AI methods were reviewed and it has been sorted. The precision of the proposed models shift contingent upon the apparatus utilized, the dataset utilized, the quantity of qualities and it is recorded in the dataset. Information investigation was improved comprehension of the dataset. Different machine learning calculations were prepared and tried on the dataset. Among every one of the calculations KNN and random forest calculation had the best exactness of 88.52%.

We presume that to fabricate an exact heart disease, dataset with adequate examples and right information should be utilized. Likewise, an appropriate calculation should be utilized when building up a forecast model. At last, the field of utilizing AI for classification of heart disease is a significant field, and it can help both medical services experts and patients. It is as yet a developing field, and in spite of the monstrous accessibility of patients informations in medical clinics or centres , very little of it is distributed.

#### 5. REFERENCES



- [1] I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," *Artificial Intelligence in Medicine*, vol. 23, no. 1, pp. 89–109, 2001.
- [2] J. Soni et al., "Intelligent and effective heart disease prediction system using weighted associative classifiers," *International Journal on Computer Science and Engineering*, vol. 3, no. 6, pp. 2385–2392, 2011.
- [3] N. Khateeb and M. Usman, "Efficient heart disease prediction system using k-nearest neighbor classification technique," in *Proceedings of the International Conference on Big Data and Internet of Thing (BDIOT)*, New York, NY, USA: ACM, 2017, pp. 21–26.
- [4] H. Almarabeh and E. Amer, "A study of data mining techniques accuracy for healthcare," *International Journal of Computer Applications*, vol. 168, no. 3, pp. 12–17, Jun 2017.
- [5] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," *Journal of Intelligent Learning Systems and Applications*, vol. 9, no. 01, pp. 1–16, 2017.
- [6] S. Pouriyeh et al., "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in *Proceedings of IEEE Symposium on Computers and Communications (ISCC)*, Heraklion, Greece: IEEE, July 2017, pp. 204–207.
- [7] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," *Expert systems with applications*, vol. 36, no. 4, pp. 7675–7680, 2009.
- [8] N. Waghulde and N. Patil, "Genetic neural approach for heart disease prediction," *International Journal of Advanced Computer Research*, vol. 4, no. 3, pp. 778, 2014.
- [9] S. Garcia et al., "Big data preprocessing: methods and prospects," *Big Data Analytics*, vol. 1, no. 1, p. 9, Nov 2016.
- [10] A. Janosi et al., "Heart disease data set," Jul 1988. [Online]. Available: [http://archive.ics.uci.edu/ml/datasets/heart Disease](http://archive.ics.uci.edu/ml/datasets/heart+Disease).
- [11] K. Vembandasamy, R. Sasipriya, and E. Deepa, "Heart diseases detection using naive bayes algorithm," *International Journal of Innovative Science, Engineering & Technology*, vol. 2, no. 9, pp. 441–444, 2015.
- [12] D. Medhekar, M. Bote, and S. Deshmukh, "Heart disease prediction system using naive bayes," *International Journal of Enhanced Research In Science Technology & Engineering*, vol. 2, no. 3, pp. 1–5, 2013.
- [13] A. Chen et al., "HDPS: Heart disease prediction system," in *Computing in Cardiology*, Hangzhou, China: IEEE, 2011, pp. 557–560.
- [14] C. Dangare and S. Apte, "A data mining approach for prediction of heart disease using neural networks," *International Journal of Computer Engineering & Technology*, vol. 3, no. 3, pp. 30–40, 2012.
- [15] V. Sabarinathan and V. Sugumaran, "Diagnosis of heart disease using decision tree," *International Journal of Research in Computer Applications & Information Technology*, vol. 2, no. 6, pp. 74–79, 2014.
- [16] J. Patel et al., "Heart disease prediction using machine learning and data mining technique," *Heart Disease*, vol. 7, no. 1, pp. 129–137, 2015.
- [17] M. Shouman, T. Turner, and R. Stocker, "Applying k-nearest neighbour in diagnosing heart disease patients," *International Journal of Information and Education Technology*, vol. 2, no. 3, pp. 220, 2012.
- [18] W. Wiharto, H. Kusnanto, and H. Herianto, "Performance analysis of multiclass support vector machine classification for diagnosis of coronary heart diseases," *International Journal on Computational Science & Applications*, vol. 5, no. 5, pp. 27–37, 2015.
- [19] S. Amin, K. Agarwal, and R. Beg, "Genetic neural network based data mining in prediction of heart disease using risk factors," in *IEEE Conference on Information Communication Technologies*. Thuckalay, Tamil Nadu, India, April 2013, pp. 1227–1231.

- [20] N. Amma, “Cardiovascular disease prediction system using genetic algorithm and neural network,” in International Conference on Computing, Communication and Applications. Dindigul, Tamilnadu, India: IEEE, Feb 2012, pp. 1–5.  
March 2008, pp. 108–115.