International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

Review of Machine Learning Methods for Heart Disease Diagnosis and Prediction

Vikas Mishra, Ram Kailash Guptas
Dept of Computer Science,
BN college of engineering and technology, Lucknow, India
vikas5misra@gmail.com

Abstract: Heart disease remains one of the leading causes of mortality worldwide. Accurate and early diagnosis plays a vital role in reducing mortality rates and enabling effective treatment. Traditional diagnostic methods rely heavily on clinical expertise and tests, which can sometimes be time-consuming and error-prone. With the increasing availability of healthcare data and advancements in computational power, machine learning (ML) techniques have emerged as powerful tools for predicting and diagnosing heart disease. This paper provides a comprehensive review of various ML methods applied to heart disease prediction, highlighting their effectiveness, strengths, limitations, and future prospects.

Keywords: Heart Disease Prediction, Machine Learning, Cardiovascular Diagnosis, Medical Data Classification. Supervised Learning, Support Vector Machine (SVM)

1. Introduction

Heart disease, also known as cardiovascular disease (CVD), encompasses a range of conditions that affect the heart and blood vessels. It is responsible for a significant percentage of global deaths each year. Early diagnosis is crucial for effective treatment and prevention. In recent years, machine learning has shown great promise in the medical field by automating and enhancing diagnostic procedures. The objective of this review is to explore the role of ML in heart disease prediction and diagnosis by examining the commonly used techniques, datasets, performance metrics, and real-world applications.

2. Machine Learning in Healthcare

Machine learning involves training algorithms on data to make predictions or decisions without being explicitly programmed. In healthcare, ML is used for tasks such as disease diagnosis, risk prediction, drug discovery, and medical image analysis. For heart disease, ML helps in identifying patterns and risk factors from patient data, thereby assisting clinicians in making informed decisions.

3. Commonly Used Datasets

Several publicly available datasets have been used in the development and evaluation of ML models for heart disease prediction. The most widely used datasets include:

- Cleveland Heart Disease Dataset: Provided by the UCI Machine Learning Repository, it includes 303 patient records with 14 attributes.
- Framingham Heart Study Dataset: A longitudinal dataset with risk factor data and outcomes.
- Statlog (Heart) Dataset: Another version from UCI, with 270 instances and 13 attributes.

4. Machine Learning Algorithms for Heart Disease Prediction

4.1 Logistic Regression (LR)

A statistical model that estimates the probability of a binary outcome. It is simple, interpretable, and often used as a baseline.

4.2 Decision Trees (DT)

Decision trees model decisions and their possible consequences. They are easy to interpret and visualize.

4.3 Support Vector Machines (SVM)

SVMs find the optimal hyperplane to separate data into classes. They are effective for high-dimensional data.

4.4 Naive Bayes (NB)

Based on Bayes' theorem, NB assumes independence among predictors. It is computationally efficient and works well with large datasets.

4.5 K-Nearest Neighbors (KNN)

KNN classifies instances based on the majority class among their k-nearest neighbors. It is simple but sensitive to noisy data.

4.6 Artificial Neural Networks (ANN)

ANNs are modeled after the human brain and are capable of capturing complex relationships in data.

4.7 Random Forest (RF) An ensemble of decision trees that improves predictive accuracy and controls overfitting.

4.8 Gradient Boosting Machines (GBM) Boosting methods like XGBoost and LightGBM have shown high performance in heart disease prediction tasks.

4.9 Deep Learning Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) are applied for more complex data and larger datasets.

5. Performance Evaluation Metrics

The following metrics are commonly used to evaluate ML models:

- Accuracy: Proportion of correctly predicted instances.
- Precision: Proportion of true positives among predicted positives.

- Recall (Sensitivity): Proportion of true positives among actual positives.
- F1 Score: Harmonic mean of precision and recall.
- AUC-ROC Curve: Measures model's ability to distinguish between classes.

6. Comparative Analysis

Many studies compare ML algorithms for heart disease prediction. Typically, ensemble methods (RF, GBM) and deep learning models outperform simpler algorithms like LR and NB in terms of accuracy and robustness. However, simpler models are preferred when interpretability is critical.

Algorithm	Accuracy (%)	Strengths	Weaknesses
Logistic Regression	83-85	Easy to implement	Limited to linear relationships
Decision Trees	80-85	Interpretable	Prone to overfitting
Random Forest	85-90	High accuracy	Less interpretable
SVM	83-88	Handles high dimensions	Requires tuning
ANN	85-92	Captures complex patterns	Requires large data
KNN	78-85	Simple	Computationally expensive

7. Challenges and Limitations

- **Data Quality**: Missing or noisy data can reduce model performance.
- **Imbalanced Data**: In some datasets, the number of positive and negative cases is skewed.
- **Model Interpretability**: Complex models like neural networks lack transparency.
- **Integration in Clinical Settings**: Adopting ML in real-world practice requires validation and regulatory approval.

8. Related

Shaikh Abdul Hannan et al. [5] used a Radial Basis Function(RBF) to predict the medical prescription for heart disease. About 300 patient's data were collected from the Sahara Hospital, Aurangabad. RBFNN (Radial Basis Function–Neural Network) can be described as a three-layer feed forward structure. The three layers are the input layer, hidden layer and output layer. The hidden layer consists of a number of RBF units (nh) and bias (bk). Each neuron on the hidden layer uses a radial basis function as a nonlinear transfer function to operate on the input data. The most often used RBF is usually a Gaussian function. Designing a RBFNN involves selecting centres, number of hidden layer units, width and weights. The various ways of selecting the centres are random subset selection, k-means clustering and others. The methodology was applied in MATLAB. Obtained results show that radial basis function can be successfully used (with an accuracy of 90 to 97%) for prescribing the medicines for heart disease.

AH Chen et al. [6] presented a heart disease prediction system that can aid doctors in predicting heart disease status based on the clinical data of patients. Thirteen important clinical features such as age, sex, chest pain type were selected. An artificial neural network algorithm was used for classifying heart disease based on these clinical features. Data was collected from machine learning repository of UCI .The artificial neural network model contained three layers i.e. the input layer, the hidden layer and the output layer having 13 neurons, 6 neurons and 2 neurons respectively. Learning Vector Quantization (LVQ) was used in this study. LVQ is a special case of an artificial neural network that applies a prototype-based supervised classification algorithm. C programming language was used as a tool to implement heart disease classification and prediction trained via artificial neural network.The system was developed in C and C# environment.The accuracy of the proposed method for prediction is near to 80%.

Mrudula Gudadhe et al.[7] presented a decision support system for heart disease classification. Support vector machine (SVM) and artificial neural network (ANN) were the two main methods used in this system. A multilayer perceptron neural network (MLPNN) with three layers was employed to develop a decision support system for the diagnosis of heart disease. This multilayer perceptron neural network was trained by back-propagation algorithm which is computationally an efficient method. Results showed that a MLPNN with back-propagation technique can be successfully used for diagnosing heart disease.

Manpreet Singh et al. [8] proposed a heart disease prediction system based on Structural Equation Modelling (SEM) and Fuzzy Cognitive Map (FCM).They used Canadian Community Health Survey (CCHS) 2012 dataset. Here, twenty significant attributes were used. SEM is used to generate the weight matrix for the FCM model which then predicts a possibility of cardiovascular diseases. A SEM model is defined with correlation between CCC 121(a variable which defines whether the respondent has heart disease) along with 20 attributes. To construct FCM a weight matrix representing the strength of the causal relationship between concepts must be constructed first. The SEM defined in the previous section is now used as the FCM though they have achieved the required ingredients (i.e. weight matrix, concepts and causality).80% of the data set was used for training the SEM model and the remaining 20% for testing the FCM model. The accuracy obtained by using this model was 74%. Carlos Ordonez [9] has studied association rule mining with the train and test concept on a dataset for heart disease prediction. Association rule mining has a disadvantage that it produces extremely large number of rules most of which are medically irrelevant. Also in general, association rules are mined on the entire data set without validation on an independent sample. In order to solve this, the author has devised an algorithm that uses search constraints to reduce the number of rules. The algorithm then searches for association rules on a training set and finally validates them on an independent test set. The medical significance of discovered rules is then evaluated with support, confidence and lift. Search constraints and test set

validation significantly reduce the number of association rules and produce a set of rules with high predictive accuracy. These rules represent valuable medical knowledge.

Prajakta Ghadge et al. [10] have worked on an intelligent heart attack prediction system using big data. Heart attack needs to be diagnosed timely and effectively because of its high prevalence. The objective of this research article is to find a prototype intelligent heart attack prediction system that uses big data and data mining modeling techniques. This system can extract hidden knowledge (patterns and relationships) associated with heart disease from a given historical heart disease database. This approach uses Hadoop which is an open-source software framework written in Java for distributed processing and storage of huge datasets. Apache Mahout produced by Apache Software Foundation provides free implementation of distributed or scalable machine learning algorithms. Record set with 13 attributes (age, sex, serum cholesterol, fasting blood sugar etc.) was obtained from the Cleveland Heart Database which is available on the web. The patterns were extracted using three techniques i.e. neural network, Naïve Bayes and Decision tree. The future scope of this system aims at giving more sophisticated prediction models, risk calculation tools and feature extraction tools for other clinical risks.

Asha Rajkumar et al. [11] worked on diagnosis of heart disease using classification based on supervised machine learning. Tanagra tool is used to classify the data, 10 fold cross validation is used to evaluate the data and the results are compared. Tanagra is a free data mining software for academic and research purposes. It suggests several data mining methods from explanatory data analysis, statistical learning, machine learning and database area. The dataset is divided into two parts, 80% data is used for training and 20% for testing. Among the three techniques, Naïve Bayes shows lower error ratio and takes the least amount of time.

K. S. Kavitha et al. [12] modelled and designed an evolutionary neural network for heart disease detection. This research describes a new system for detection of heart diseases using feed forward neural architecture and genetic algorithm. The proposed system aims at providing easier, cost effective and reliable diagnosis for heart disease. The dataset is obtained from UCI repository. The weights of the nodes for the artificial neural network with 13 input nodes, 2 hidden nodes and 1 output node are once set with gradient descent algorithm and then with genetic algorithm. The performances of these methods are compared and it is concluded that genetic algorithm can efficiently select the optimal set of weights. In genetic algorithm tournament selection is a method of selecting an individual from a population of individuals. This work finds that more members are coming from the offspring population. It is an indication for generation of fitter offsprings which leads to greater diversity and exploration of search space. With the help of this work, expert disease prediction systems can be developed in the future.

K. Sudhakar et al. [13] studied heart disease prediction using data mining. The data generated by the healthcare industry is huge and "information rich". As such, it cannot be interpreted manually. Data mining can be effectively used to predict diseases from these datasets. In this paper, different data mining techniques are analyzed on heart disease database. Classification techniques such as Decision tree, Naïve Bayes and neural network are applied here .Associative classification is a new and efficient technique which integrates association rule mining and classification to a model for prediction and achieves maximum accuracy. In conclusion, this paper analyzes and compares how different classification algorithms work on a heart disease database.

Shantakumar B. Patil et al. [14] obtained important patterns from heart disease database for heart attack prediction. Enormous amount of data collected by the healthcare industry is unfortunately not 'mined' properly to find concealed information that can predict heart attack. Here, the authors have proposed MAFIA algorithm (Maximal Frequent Itemset Algorithm) to do so using Java. The data is preprocessed first, and then clustered using k-means algorithm into two clusters and the cluster significant to heart attack is obtained. Then frequent patterns are mined from the item set and significance weightages of the frequent data are calculated. Based on these weightages of the attributes(ex- age, blood pressure, cholesterol and many others), patterns significant to heart attack are chosen. This pattern can be further used to develop heart attack prediction systems.

Sairabi H. Mujawar et al. [15] predicted heart disease using modified k-means and Naïve Bayes. Diagnosis of heart disease is a complex task and requires great skills. The dataset is obtained from Cleveland Heart Disease Database. The attribute "Disease" with a value '1' indicates the presence of heart disease and a value '0' indicates the absence of heart disease. Modified k-means works on both categorical and combinational data which we encounter here. Using two initial centroids we obtain two farthest clusters. It finally gives a suitable number of clusters. Naive Bayes's creates a model with predictive capabilities. This predictor defines the class to which a particular tuple should belong to. This predictor has 93 % accuracy in predicting a heart disease and 89% accuracy in cases where it detected that a patient doesn't have a heart disease.

S. Suganya et al. [16] predicted heart disease using fuzzy cart algorithm. Fuzziness was introduced in the measured data to remove the uncertainty in data. A membership function was thus incorporated. Minimum distance CART classifier was used which proved efficient with respect to other classifiers of parametric techniques. The heart disease dataset is initially segregated into attributes that increase heart disease risk. Then fuzzy membership function is applied to remove uncertainty and finally ID-3 algorithm is run recursively through the non-leaf branches until all the data have been classified. The proposed method is implemented in Java.

Ashwini Shetty A et al. [17] proposed different data mining approaches for predicting heart disease. Their research work analyses the neural network and genetic algorithm to predict heart diseases. The initial weight of the neural network is found using genetic algorithm which is the main advantage of this method. Here, the neural network uses 13 input layers, 10 hidden layers and 2 output layers. The inputs are the attribute layers (here 13 attributes are used namely age, resting heart rate, blood pressure, blood sugar and others). Levenberg-Marquardt back propagation algorithm is used for training and testing. Optimization Toolbox is used to implement this system. 'configure' function is used with neural network where each weight lies between -2 to 2.Fitness function that

is being used in the genetic algorithm is the Mean Square Error (MSE). Genetic algorithm is used for adjustment of weights. Based on MSE, fitness function will be calculated for each chromosome. Once selection is done, crossover and mutation in genetic algorithm replaces the chromosome having lower adaption with the better values. Fitter strings are obtained by optimizing the solution which corresponds to interconnecting weights and threshold of neural network. The resulting lower values those are close to zero, represent the generalized format of the network which is ready for classification problem. The system calculates accuracy using MATLAB. Preprocessing is done using WEKA. The results show that the hybrid system of genetic algorithm and neural network works much better than the performance of neural network alone.

9. Conclusion

Machine learning methods offer significant promise in improving the accuracy and efficiency of heart disease diagnosis and prediction. While challenges remain in terms of data quality, interpretability, and clinical adoption, ongoing advancements continue to bridge the gap between algorithm development and real-world implementation. ML has the potential to transform cardiovascular healthcare by enabling earlier diagnosis, personalized treatment, and better patient outcomes.

References

[1] Nidhi Bhatla, and Kiran Jyoti, Oct. 2012,"An Analysis of Heart Disease Prediction using Different Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT), Vol. 1, Issue 8, ISSN: 2278-0181, pp. 1-4.

[2] Sumitra Sangwan, and Tazeem Ahmad Khan, Mar. 2015,"Review Paper Automatic Console for Disease Prediction using Integrated Module of A-priori and k-mean through ECG Signal",International Journal For Technological Research In Engineering, Vol. 2, Issue 7, ISSN(Online): 2347-4718, pp. 1368-1372.

[3] Rishi Dubey, and Santosh Chandrakar, Aug. 2015,"Review on Hybrid Data Mining Techniques for The Diagnosis of Heart Diseases in Medical Ground" ,Vol. 5, Issue 8, ISSN: 2249-555X, pp. 715-718.

[4] Ashish Chhabbi,Lakhan Ahuja,Sahil Ahir, and Y. K. Sharma,19 March 2016,"Heart Disease Prediction Using Data Mining Techniques", International Journal of Research in Advent Technology,E-ISSN:2321-9637,Special Issue National Conference "NCPC-2016", pp. 104-106.

[5] Shaikh Abdul Hannan, A.V. Mane, R. R. Manza,and R. J. Ramteke, Dec 2010,"Prediction of Heart Disease Medical Prescription using Radial Basis Function",IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), DOI: 10.1109/ICCIC.2010.5705900 ,28-29 .

[6] AH Chen, SY Huang, PS Hong, CH Cheng, and EJ Lin,2011, "HDPS: Heart Disease Prediction System",Computing in Cardiology, ISSN: 0276-6574, pp.557- 560. [7] Mrudula Gudadhe, Kapil Wankhade, and Snehlata Dongre, Sept 2010,"Decision Support System for Heart Disease Based on Support Vector Machine and Artificial Neural Network",International Conference on Computer and Communication Technology (ICCCT),DOI:10.1109/ICCCT.2010.5640377, 17-19.

[8] Manpreet Singh, Levi Monteiro Martins, Patrick Joanis, and Vijay K. Mago,2016, "Building a Cardiovascular Disease Predictive Model using Structural Equation Model & Fuzzy Cognitive Map",IEEE International Conference on Fuzzy Systems (FUZZ),pp. 1377-1382.

[9] Carlos Ordonez,2006, "Association Rule Discovery With the Train and Test Approach for Heart Disease Prediction", IEEE Transactions on Information Technology in Biomedicine (TITB), pp. 334-343, vol. 10, no. 2.

[10] Prajakta Ghadge,Vrushali Girme, Kajal Kokane, and Prajakta Deshmukh, 2016,"Intelligent Heart Attack Prediction System Using Big Data", International Journal of Recent Research in Mathematics Computer Science and Information Technology,Vol. 2, Issue 2, pp.73-77, October 2015–March.

[11] Asha Rajkumar, and Mrs G. Sophia Reena, 2010, "Diagnosis of Heart Disease using Data Mining Algorithms",Global Journal of Computer Science and Technology,Vol. 10,Issue 10, pp.38-43, September.

[12] K. S. Kavitha, K. V. Ramakrishnan, and Manoj Kumar Singh, September 2010,"Modelling and Design of Evolutionary Neural Network for Heart Disease Detection" ,International Journal of Computer Science Issues (IJCSI), Vol. 7, Issue 5, pp. 272-283.

[13] K.Sudhakar, and Dr. M. Manimekalai, January 2014, "Study of Heart Disease Prediction using Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 1,pp. 1157-1160.

[14] Shantakumar B. Patil,and Dr. Y. S. Kumaraswamy, February 2009, "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction", IJCSNS International Journal of Computer Science and Network Security, Vol. 9 , No. 2,pp. 228-235.

[15] Sairabi H. Mujawar, and P. R. Devale, October 2015,"Prediction of Heart Disease using Modified k-means and by using Naive Bayes", International Journal of Innovative Research in Computer and Communication Engineering(An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 10, pp. 10265-10273.

[16] S. Suganya, and P.Tamije Selvy, January 2016, "A Proficient Heart Disease Prediction Method using Fuzzy-Cart Algorithm", International Journal of Scientific Engineering and Applied Science (IJSEAS), Vol. 2, Issue1,ISSN: 2395-3470.

[17] Ashwini Shetty A, and Chandra Naik, May 2016,"Different Data Mining Approaches for Predicting Heart Disease", International Journal of Innovative Research in Science,Engineering and Technology(An ISO 3297: 2007 Certified Organization), Vol. 5, Special Issue 9, pp. 277-281.

[18] K Cinetha, and Dr. P. Uma Maheswari, Mar.-Apr. 2014,"Decision Support System for Precluding Coronary Heart Disease using Fuzzy Logic.", International Journal of Computer Science Trends and Technology (IJCST), Vol. 2, Issue 2, pp. 102-107.

[19] Indira S. Fal Dessai,2013,"Intelligent Heart Disease Prediction System Using Probabilistic Neural Network", International Journal on Advanced Computer Theory and Engineering (IJACTE),Vol. 2, Issue 3, pp. 38-44.

[20] Mai Shouman, Tim Turner, and Rob Stocker, June 2012,"Applying k-Nearest Neighbors in Diagnosing HeartDiseasePatients",International Journal of Information and Education Technology, Vol. 2, No. 3,pp. 220-223.
[21] Serdar AYDIN, Meysam Ahanpanjeh,and Sogol Mohabbatiyan,February 2016, "Comparison And Evaluation of Data Mining Techniques in the Diagnosis of Heart Disease",International Journal on Computational Science & Applications (IJCSA), Vol. 6,No.1, pp. 1-15.
[22] G. Purusothaman, and P. Krishnakumari, June 2015,"A Survey of Data Mining Techniques on Risk Prediction: Heart Disease", Indian Journal of Science and Technology, Vol. 8(12), DOI:10.17485/ijst/2015/v8i12/58385, pp. 1-5.
[23] Deepali Chandna, 2014,"Diagnosis of Heart Disease Using Data Mining Algorithm", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 5 (2), pp. 1678-1680.
[24] S. Prabhavathi, and D. M. Chitra, Jan. 2016, "Analysis and Prediction of Various Heart Diseases using DNFS Techniques", International Journal of Innovations in Scientific and Engineering Research(IJISER), Vol.2, Issue 1, pp. 1-7.
[25] Jaymin Patel, Prof. Tejal Upadhyay, and Dr. Samir Patel, Sep 2015-Mar 2016, "Heart Disease Prediction using Machine Learning and Data Mining Technique", Vol. 7, No.1, pp. 129-137.
[26] Vikas Chaurasia, and Saurabh Pal, 2013, "Early Prediction of Heart Diseases Using DataMiningTechniques",Caribbean Journal of Science and Technology, ISSN: 0799-3757, Vol.1, pp. 208-218.
[27] Gunsai Pooja Dineshgar, and Mrs. Lolita Singh, February 2016, "A Review on Data Mining for Heart Disease Prediction",International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE), Vol. 5, Issue 2, pp. 462-466.
[28] Anurag et. al., "Load Forecasting by using ANFIS", International Journal of Research and Development in Applied Science and Engineering, Volume 20, Issue 1, 2020
[29] Raghawend, Anurag, "Detect Skin Defects by Modern Image Segmentation Approach, Volume 20, Issue 1, 2020
[30] Sharan Monica L, and Sathees Kumar B,February 2016, "Analysis of Cardiovascular Heart Disease Prediction Using Data Mining Techniques", International Journal of Modern Computer Science (IJMCS), ISSN: 2320-7868(Online), Vol. 4, Issue 1, pp. 55-58.
[31] UCI Machine Learning Repository: Heart Disease Dataset.
[32] Detrano, R., et al. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*.
[33] Rajkomar, A., et al. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347-1358.
[34] Deo, R. C. (2015). Machine learning in medicine. *Circulation*, 132(20), 1920–1930.
[35] Krittanawong, C., et al. (2017). Machine learning prediction in cardiovascular diseases: A meta-analysis. *Scientific Reports*, 7, 3954.