

Heart Disease Prediction Methods

Mansi Gunsai, Shreya Patel, Kinjal V. Joshi

Abstract: In recent times, heart diseases are considered one of the deadliest causes of mortality and morbidity among the population of the world. Predicting the probability of the occurrence of cardiovascular diseases has become one of the most important objectives of the medical analysis system. The conventional methods have proved to be inefficient in prior prediction of heart diseases because of several contributing risk factors like diabetes, high blood pressure, high cholesterol, abnormal pulse rate, and others. Due to such limitations, medical practitioners rely on various modern Machine Learning and Data Mining approaches such as ANN, Naïve Bayes, SVM, etc. Such models have proved to be effective in providing accurate predictions from the huge amount of medical data available. The main aim of this paper is to analyse various machine learning approaches adopted in different research works and to deduce which techniques are most beneficial and precise.

Keywords: Heart Disease Prediction, Machine Learning, Artificial Neural Network, Naïve Bayes, Decision Tree, Genetic Algorithm

I. INTRODUCTION

Heart disease is one of the leading causes of death among men and women of all ages in almost the whole world. Over 17 million deaths have been estimated annually due to Cardiovascular Diseases in the past few years, which represents about 31% of all deaths globally and out of these, 85% are caused by heart attacks and strokes. According to the statistics of the World Health Organisation and Centre for Disease Control and Protection, one in every four people dies due to heart diseases. The number of prevalent cases of cardiovascular diseases is increasing day by day. More than half of the total deaths were in people younger than 70. Moreover, one in five heart attacks is silent-damage: the damage is done but the person is not aware of it. And thus, it has become crucial to prevent premature deaths from heart diseases.

The general term heart disease actually comprises a range of conditions that affect the heart. It includes diseases of the blood vessels, coronary artery disease (CAD) and peripheral artery disease (PAD) abnormal heart beats, or arrhythmias or congenital heart disease, and the thickening or enlarging of the heart, a condition called cardiomyopathy. There are a number of causes for heart diseases also depending on the type of disease. But the most critical ones are high blood pressure, high cholesterol and diabetes. Also, excessive consumption of tobacco (smoking), caffeine, alcohol or narcotic drugs can lead to the dangers of heart attacks. Other external factors such as lack of physical inactivity, obesity, unhealthy diet can cause heart diseases. Even a family history of heart disease can be one of the reasons.

Revised Manuscript Received on May 20, 2020.

Mansi Gunsai Department of Computer Engineering from G H Patel College of Engineering & Technology, Anand, Gujarat, India.

Shreya Patel, Department of Computer Engineering from G H Patel College of Engineering & Technology, Anand, Gujarat, India.

Prof Kinjal Joshi, Department of Computer Engineering from G H Patel College of Engineering & Technology, Anand, Gujarat, India.

Medical Research Systems are developed rapidly to avoid the risk of heart diseases. Medical Diagnosis Systems which are primarily based on various machine learning techniques are used for predicting heart diseases. Such techniques transform large collections of raw healthcare data into information that can help to make informed decisions and predictions.

II. LITERATURE SURVEY

Traditional way of predicting Heart disease is a doctor's examination or number of medical tests such as ECG, Stress Test, and Heart MRI etc. But, nowadays there is a huge amount of data available of various medical tests and examinations in the health-care industry so it is possible to build a Heart disease prediction system. The goal of this survey is to determine the machine learning technique with maximum accuracy. The data source used for most of the research papers we surveyed on is the Cleveland Heart Disease Dataset from UCI Machine Learning Repository.

Table 2. Dataset Description

Attributes	Units	Description
Age	Numeric value in years	Age of the patients
Sex	1 = male 2 = female	Gender of the patient
cp	1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 = asymptotic	Type of chest pain experienced by the patient
trestbps	Numeric value in mmHg	Resting blood pressure value of the patient
chol	Numeric value in mg/dl	Serum Cholesterol
fb	1 = fasting blood sugar > 120mg/dl 0 = otherwise	Fasting blood sugar Compares fbs value of the patient with 120 mg/dl
restecg	0 = normal 1 = having ST-T wave abnormality 2 = left ventricular hypertrophy	Resting electrocardiographic results
Thalach	Numeric value in bpm (beats per minute)	Maximum heart rate achieved by the patient
exang	1 = yes 0 = no	Exercise induced angina
oldpeak	Numeric Value	ST depression induced by exercise relative to rest
slope	1 = upsloping 2 = flat 3 = downsloping	Peak exercise ST segment

Heart Disease Prediction Methods

Ca	Numeric Value	Number of major vessels (0-3) coloured by fluoroscopy
thal	3 = normal 6 = fixed defect 7 = reversible defect	Thalassemia
Target	0 = absence 1, 2, 3, 4 = present	Diagnosis of the Heart Disease

2.1 Technique based on Genetic algorithm hybrid with ANN

Ankita Dewan and Meghna Sharma et al. [1] has used various data mining techniques that can solve complicated queries and thus helps medical assist doctors to make smart decisions. The dataset is taken from UCI repository which includes 13 attributes such as sex, Blood pressure, cholesterol and added 2 more attributes obesity and smoking. Here, they have proposed a Backpropagation technique using an Artificial neural network. They also proposed another technique that is using the Genetic Algorithm as an optimizer due to the drawback of Backpropagation Algorithm of being stuck in local minima. The proposed methodology can be applied in the future with the intention of getting 100% accuracy on minimal error.

2.2 CANFIS model integrated with Genetic Algorithm

LathaParthiban and R. Subramanian et al. [2] proposed Intelligent Heart Disease Prediction System using CANFIS model and Genetic Algorithm in 2008. The CANFIS model integrates fuzzy inputs with a modular neural network to rapidly and accurately approximate complex functions. Here they use fuzzy inference which consists of membership functions (explanatory set of rules) with the power of neural networks which helps to solve complex problems more efficiently than in neural networks. To improve the learning of CANFIS they used Genetic algorithms that search the best number of MF for each input and optimization of control parameters like learning rate and momentum coefficient. It has been shown here that GA is a very powerful technique for auto-tuning of CANFIS parameters and selection of optimized feature set.

2.3 PSO (Particle Swarm Optimization) along with Neural Network Technique

Majid and Omid et al. [2] have done their research work in one of the challenges of the medical field. Here, they have used classification and feature selection approaches. They have proposed the method PSO with (FFBP + feature ranking) which reduces the cost by reducing the 13 features set to 8 features set which are (Age, Sex, Chol, FBS, RestECG, Exang, Oldpeak, Slope). In the first stage of this work as the dataset contained the data of healthy as well as sick people so they divided into a group of healthy and sick people. Then in the second stage, 8192 subsets were extracted from the total dataset. Then in the third stage with the help of PSO algorithm a learning classifier is assigned to all the subsets to find the subset with the highest accuracy and with the lowest cost and time. Four classifier methods are used in this approach SMO, C4.5, MLP, and FFBP and the results section V suggests that the use of FFBP neural network and feature selection using PSO algorithm is more efficient. Due to feature ranking accuracy have been

improved by 2.38% and the accuracy is 91.94%, reduction in cost is also due to feature selection and feature ranking.

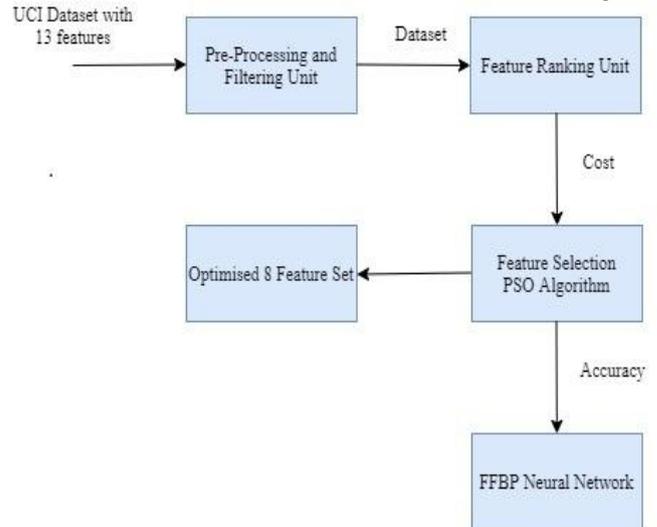


Fig. 1. Applying PSO for heart disease prediction^[2]

2.4 Binary cuckoo optimization algorithm (BCOA) with SVM model

SadafRoostae& Hamid Reza Ghaffary et al. [4] has used the UCI dataset to evaluate the heart attack. They have proposed SVM for model construction and binary cuckoo optimization algorithm (BCOA) is used for feature selection. The final accuracy of the model is 84.44%, sensitivity is 86.49% and specificity is 81.49%. The specificity measure is lower than sensitivity measure which means that the model is more powerful in diagnosing patients as compared to healthy people.

2.5Data mining extension query language for model building

SellappanPalaniappan, RafiahAwang et al. [5] has used the Cleveland Heart disease database a total of 909 records with 15 medical attributes and developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naïve Bayes and Neural Network. IHDPS can answer complex “what if” queries which traditional decision support systems cannot. Data mining extension query language is used to build the models. Classification matrix and Life chart are used for evaluating the effectiveness of the models. Here they figured out that on this database Naïve Bayes to be the most effective with an accuracy of 86.53% with the highest prediction of heart disease patients and decision tree to be with the accuracy of 89% and neural network with 85%. However, all three models can answer complex queries. Detailed patient’s profile is available in decision tree but Naïve Bayes has fared better than a decision tree as it identifies all noteworthy medical predictors. They can further expand this system by using other attributes instead of using only 15 attributes that they have used here and also accuracy could be improved by using time-series, clustering and association rules.

2.6 Using Artificial feed forward neural network

NouraAjam et al. [6] is making use of the Artificial Neural Network in the diagnosis of Heart Disease. The neural network is trained with a heart disease database by using a typical feed forward neural network model and back-propagation learning algorithm. He used the Feed-forward Neural Network as a classifier to detect between the absence or presence of disease. The model consisted of an input layer of 13 neurons as the dataset was of 13 attributes, the hidden layer consisted of 20 neurons and an output layer with only 1 neuron. In this paper, the mean squared error (MSE) is equal to 0.1071 and classification accuracy is 88%.

2.7 Algorithm based on Association rules

Association rule mining is used to find associations or correlations among the item sets. So, Association rules have great potential to improve disease prediction in the medical domain. The use of association rules on medical data records with heart perfusion measurements is explored for the first time in [15]. The problem of identifying constrained association rules for heart disease prediction was studied by Carlos Ordonez [16]. In [2] paper by M. Akhiljabbbar, Dr. Priti Chandra they used Gini index as feature selection method and filter rules further by applying Z-Statistics and genetic algorithm. And using Apriori algorithm for classification with an accuracy of 98%.

Table 1. Performance analysis of Heart disease prediction using Machine learning techniques

Author	Techniques Used	Dataset	Accuracy
Khemphila&Bonjing (2011)	Feature selection: Information Gain Classification: Multi-Layer Perceptron (MLP) with Backpropagation learning algorithm	Cleveland heart disease data (UCI Repository)	Accuracy: Training:89.56% Testing:80.99%
Akhil, Dr. Priti& Dr. Deekshatulu (2012)	Feature selection: Gini Index Classification: Apriori algorithm	Heart disease data (Andhra Pradesh)	Accuracy:98%
Syed Umar Amin, Kavita Agarwal (2013)	Classification: Neural Network & Genetic Algorithm	American Heart Association	Accuracy: Training:96.2% Testing:92%
R. Chitra and Dr.V. Seenivasagam (2013)	Classification: Artificial Neural Network with back	UCI Repository (University of Californ	Accuracy with ANN: Training:72.6% Testing:79.45%

	propagation network cascaded correlation Neural Network	ia, Irvine C.A)	Accuracy with CNN: Training:78.55% Testing:85%
NouraAjam (2015)	Classification: Artificial Neural Network	Cleveland heart disease data (UCI Repository)	Accuracy:88%
Majid Feshki&OmidS oजूodi (2016)	Feature selection & ranking: PSO algorithm Classification: FFBP Neural Network	Cleveland heart disease data (UCI Repository)	Accuracy:91.94%
TülayKarayilan and Özkan (2017)	Classification: Multilayer Perceptron Neural Network with Backpropagation Algorithm	Cleveland heart disease data (UCI Repository)	Accuracy:95%
Reddy Prasad, Pidaparathi Anjali, S. Adil, N. Deepa (2019)	Classification: Logistic Regression algorithm & Naïve Bayes	UCI machine learning repository	Accuracy with Naïve Bayes:86% Accuracy with logistic Regression:86.89%

III. DISCUSSION AND FUTURE DIRECTION

The models can further be made efficient to decrease the computational requirements, save the required size of the repository, make health checklist more economical and decrease the number of parameters gathered from patients. Also, accuracy can be increased by including various other related medical attributes. Instead of categorical data, continuous data can be used with various time series data mining algorithms. Text Mining techniques for mining a huge amount of unstructured data provided by the healthcare databases can be integrated with data mining to enhance accuracy and reduce complexity. Furthermore, supplementary complex hybrid models can be utilized for more accurate predictions developed by integrating data mining and machine learning algorithms with text mining of unstructured medical data obtained from the medical institutes in huge quantities.



The efficiency can be further improved by using Genetic Algorithm which is based on subset selection process that selects the relevant input features. Optimized and relevant models providing precise predictions can be obtained by using Genetic Algorithm. For methodologies such as Decision tree or Naive Bayesian classification, the association search can be simplified by relaxing the constraints by filtering out the high number rules in the validation phase itself. Assuming that the confidence is the most important metric for validating rules, provided a smaller size dataset as ours more rigorous techniques like cross validation cannot be applied. They can only be used for larger datasets. Attributes related to artery can be used with fuzzy discretization so that rules with higher confidence can be obtained. It can be made possible to maintain a hierarchy of perfusion measurements in order to limit the rule discovery process. This could help increase or decrease sensitivity for identifying ill patients with high accuracy. A better approach to feature selection can be that the features are selected in a way that the order of choosing is defined explicitly rather than based on accuracy and costs: in terms of accessibility and testing time.

IV. CONCLUSIONS

The main objective of this paper is to analyze the various machine learning and data mining techniques adopted in different research works and infer the most accurate or beneficial technique. From the study of various research papers, we have found that in most research works the dataset used is taken from the online Cleveland database of UCI repository. The method based on Gini index as feature selection with Apriori algorithm gives 98% accuracy, on the dataset of heart disease patients of Andhra Pradesh. and the method Multilayer Perceptron Neural Network gives maximum accuracy of 95% on Cleveland dataset. It is also been found that the problem of the backpropagation technique being stuck in local minima can be solved by using optimization techniques of Genetic algorithms. From the survey, we have come up with the following observation (1) In almost all the works the small dataset is used, so for proper and accurate results, real data of heart disease patients from reputed medical institutes should be taken. (2) Hybrid techniques give more accuracy than pure machine learning techniques so we can have accuracy near to 100% using hybrid techniques.

REFERENCES

1. Ankita Dewan and Meghna Sharma, "Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification Published in: 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom).
2. M. Akhil and Dr. PritiChandrab, "Heart Disease Prediction System using Associative Classification and Genetic Algorithm "International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT, 2012.
3. Majid Feshki&OmidSojoodi, "Improving the Heart Disease Diagnosis by Evolutionary Algorithm of PSO and Feed Forward Neural Network "Published in: 2016 Artificial Intelligence and Robotics (IRANOPEN).
4. SadafRoostae& Hamid Reza Ghaffary, "Diagnosis of Heart Disease Based on Meta Heuristic Algorithms and Clustering Methods", *Journal of Electrical and Computer Engineering Innovations JECEI*, Vol. 4, No. 2, 2016.
5. SellappanPalaniappan, RafiahAwang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", Published in: 2008 IEEE/ACS International Conference on Computer Systems and Applications.

6. NouraAjam,"Heart Diseases Diagnosis using Artificial Neural Network "Network and Complex Systems ISSN 2224-610X ISSN 2225-0603 Vol.5, No.4, 2015.
7. TülayKarayilan and Özkan, "Prediction of Heart Disease Using Neural Network", Published in: 2017 *International Conference on Computer Science and Engineering (UBMK)*.
8. Reddy Prasad, Pidaparathi Anjali, S. Adil, "Heart Disease Prediction using Logistic Regression Algorithm using Machine Learning", *International Journal of Engineering and Advanced Technology (IJEAT)*ISSN: 2249 – 8958, Volume-8, Issue-3S, February 2019.
9. LathaParthiban and R. Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", *International Journal of Biological and Life Sciences* 3:3 2007.
10. Syed Umar Amin, Kavita Agarwal, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors",*Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013)*.
11. R. Chitra and Dr.V. Seenivasagam, "Heart Disease Prediction System Using Supervised Learning Classifier", *Bonfring International Journal of Software Engineering and Soft Computing*, Vol. 3, No. 1, March 2013.
12. Khempila&Boonjing, "Heart disease Classification using Neural Network and Feature Selection"2011 21st *International Conference on Systems Engineering (ICSEng)*.
13. AH Chen, SY Huang, PS Hong,"HDPS: Heart Disease Prediction System", Published in: 2011 *Computing in Cardiology*.
14. J. Thomas, "Human Heart Disease Prediction System using Data Mining Techniques",*2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT]*.
15. D. Cooke, C. Ordonez, E. V. Garcia, E. Omiecinski, E. Krawczynska,R. Folks, C. Santana, L. de Braal, and N. Ezquerra, "Data mining of large myocardial perfusion SPECT (MPS) databases to improve diagnostic decision making," *J. Nucl. Med.*, vol. 40, no. 5, 1999.
16. Carlos Ordonez," Association Rule Discovery with the Train and Test Approach for Heart Disease Prediction", *IEEE transactions on information technology in biomedicine*, vol. 10, no. 2, april 2006

AUTHORS PROFILE



Mansi Gunsais pursuing her Bachelors in Engineering in Computer Engineering from G H Patel College of Engineering & Technology, Anand, Gujarat, India and will complete graduation in May, 2020.



ShreyaPatelis pursuing her Bachelors in Engineering in Computer Engineering from G H Patel College of Engineering & Technology, Anand, Gujarat, India and will complete graduation in May, 2020. She has been the Chairperson of Computer Society, IEEE GCET SB and has served as Content Lead for Compute Newsletter Edition'7.



Prof Kinjal Joshi joined GCET in 2007. She obtained M.E. in Computer Engineering from Sardar Patel University and B.E. In Computer Engineering from Hemchandracharya North Gujarat University. She is currently pursuing PhD from Gujarat Technological University in the area of Computer Vision. Her research interests include Image Processing, Computer Vision and Machine Learning. She has published several research papers in reputed international journals and conferences. She is involved in teaching subjects at UG and PG level.