International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

A Review on Prediction of Heart Disease based on Machine Learning and Datamining Techniques

Durga Bhavani Adla[1]   Pachipala Yellamma[2]

[1]Research Scholar, Department of CSE, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Andhra Pradesh, India.
[2] Associate Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Andhra Pradesh, India.
Email: durgabhavaniphd@gmail.com

ABSTRACT
Heart is the important organ in human body which supplies blood to all organs of the body. The abnormal situation of heart is considered as heart disease. According to WHO data cardiovascular, respiratory and neonatal conditions are the top three causes of Deaths in the World. In the year 2019 Heart diseases occupies 16%(9 million) of overall deaths happened in World. From two decades there is 4 times increase in the deaths with heart diseases this is because of change in life style, lack of physical activity, food habits, obesity ,stress, cholesterol ,high blood pressure and diabetes.so there is a need to work on prediction of heart diseases to save many lives because prediction is the only way to prevent the disease. In this paper we will discuss about existing algorithms and existing work done in different machine learning and datamining techniques, which are concentrated more on the classification and prediction. main objective is to evaluate the performance of these algorithms and identify the most accurate and efficient approach for diagnosing heart diseases.
Some of the machine learning and data mining techniques are Artificial Neural Network(ANN),Decision Tree, Naive Bayes, SVM(Support Vector Machine),k-Nearest Neighbours (KNN),J48,SMO,Random forest and classification Tree.

Keywords-Machine learning; Prediction of Heart disease; Random Forest ;Decision Tree; Naive Bayes; SVM;

1.INTRODUCTION

Heart disease prediction grab the attention of many researchers in this decade because it is the main cause for most of the deaths around the world, so an early detection of the disease is needed. Some of the risk factors of heart diseases are changes in lifestyle, smoking and drinking habits, food habits, age, obesity, stress, cholesterol levels and high blood pressure.The diagnosis of heart disease is challenging and expensive task, if the heart condition of the person is known then only suitable treatment is provided, misdiagnosis leads to the death of the person. In the diagnosis of heart disease doctors first check for the signs and symptoms after that they will do physical examination of the person. Medical tests needed for heart disease treatment are like Blood test to know how much heart muscles are damaged, ECG to know the change in heartbeat, treadmill test to know how heart is working while doing some activity, echocardiogram to check heart valves and chambers problems, angiogram to know how much coronary arteries are blocked and MRI to know the structure of heart to identify the problems in detail [1].

Heart diseases remain a major global health concern, and timely prediction and classification are vital for effective intervention and treatment. Machine learning techniques have shown great promise in analyzing large datasets and identifying patterns that can aid in the diagnosis and prognosis of heart diseases. In this study, we explore and compare various machine learning algorithms to determine their effectiveness in predicting and classifying heart diseases.

The next sections of this paper consists of Literature Survey, Conclusion and Future Enhancement ,and References.

2. LITERATURE SURVEY

There are many existing works related to disease prediction like brain tumor, diabetics and heart disease using machine learning and datamining techniques. some of the works which provided good accuracy and precision are discussed here.

Author in [2] M. J. A. Junaid et al, used hybrid techniques for prediction of heart disease with the data science application, they used three algorithms Naive Bayes,Artificial neural networks and SVM and acquired the accuracy of 88% with less specificity. This hybrid model is developed based on the Cleveland dataset downloaded from UCI repository with 13 attributes.

In [3],S. Mohan et al,used hybrid technique of two machine learning algorithms that is hybrid random forest with a linear model and got the accuracy of 88.7%,this hybrid model is developed based on the Cleveland dataset downloaded from UCI repository and it can be further classified into 8 datasets and results are compared with existing algorithms, this model also got good precision and sensitivity.

In [4] C. S. Prakash et al, used only two algorithms SVM for classification and logistic regression. In this paper authors taken reduced set of attributes they used only 4 attributes age,sex,chest pain and fasting blood sugar to predict heart disease from the existing 13 attributes of Cleveland dataset from UCI repository and got good precision value for logistic regression.

In [5] C. -H. Lin et al, authors used convolution neural network on Cleveland dataset by using deep learning algorithm it works on categorical and noncategorical data and provided good accuracy. If three hidden layers are used it provides approximately 78% accuracy only with <20 neurons, less than 77% if neurons>100 is there.

In [6] N. L. Fitriyani et al, proposed HDPM through SMOTE and ENN algorithm,used two different datasets Cleveland and Statlog. This work acts as a Decision making tool for clinical support. They used strong pre-processing model including feature selection ,outlier detection and removal, clustering through DBSCAN. XGBoost algorithm is applied and got highest accuracy of 98% for Cleveland dataset and 95.9% for Statlog dataset.

In [7] Marjia Sultana et al, used many algorithms like j48,SMO,k-star,bayes net and multilayer perceptron using WEKA tool, among all these SMO got good accuracy of 89% but it is not enough to predict heart disease perfectly further improvement of accuracy is needed.

In[8] Dr. S. Seema Shedole et al, concentrated on prediction of chronic diseases they used techniques like Naive bayes, SVM and decision tree among these SVM provided good accuracy of 94% for heart disease and Naive bayes provided better accuracy of 73.58% for diabetics.

In [9]Ashok kumar Dwivedi et al, used different techniques of machine learning like Naive bayes, logistic regression ,KNN and classification Tree. among all these logistic regression provides best accuracy of 85% in their work.

Megha Shahi et al, in [10] recommended the techniques of data mining as SVM, Naïve Bayes, Association rule, KNN, ANN, and Decision Tree. The authors this paper recommended SVM as effective and provides more accuracy then other techniques.

In [11],R. Sharmila et al, proposed to use non-linear classification algorithm for heart disease prediction. They proposed to use bigdata tools such as Hadoop Distributed File System, In this paper SVM is used in parallel fashion it provides more accuracy of 85% then the sequential SVM that is of 82.3%.

Ashwini Shetty A et al, in [12] ,used Different Data Mining Approaches for Predicting Heart Disease. With WEKA tool and MATLAB. Accuracy of Neural Network 84% and Accuracy of Hybrid Systems 89%.

In most of these studies Cleveland Dataset is used which consists of 76 attributes but most of the authors used only subset of 13 attributes. This Dataset consists of 303 entries in which 137 entries are of the persons with heart disease and 160 entries are of the persons without heart disease. Remaining 6 entries are of missing values some authors considered this by taking mean value and some authors removed it by not considering. The sample dataset is given in Table 1[13] and comparision and summarization of different studies shown in Table 2.

Age	gender	chest pain	Ca:number of major vessels colored by flouroscopy	Cholestrol	Exang: Exercise induced angina or not	Fasting Blood Sugar	Restecg	Slope:the slope of the peak exercise ST segment	oldpeak:ST depression induced by exercise relative to rest	Thalach:Max heart rate Achieved	Trestbps	thalassemia 14 num 0: < 50% diameter narrowing 1: > 50% diameter narrowing diagnosis of h	Result
69	female	0	1	234	0	TRUE	2	1	0.1	131	160	0	0
69	Male	0	2	239	0	FALSE	0	0	1.8	151	140	0	0
66	Male	0	0	226	0	FALSE	0	2	2.6	114	150	0	0

65	female	0	1	282	0	TRUE	2	1	1.4	174	138	0	1
64	female	0	0	211	1	FALSE	2	1	1.8	144	110	0	0
64	female	0	0	227	0	FALSE	2	1	0.6	155	170	2	0
67	Male	2	1	277	0	FALSE	0	0	0	172	152	0	0
66	Male	2	1	278	0	FALSE	2	1	0	152	146	0	0
60	Male	0	0	240	0	FALSE	0	0	0.9	171	150	0	0
59	female	z0	0	270	0	FALSE	2	2	4.2	145	178	2	0
59	female	0	0	288	0	FALSE	2	1	0.2	159	170	2	1
53	female	2	3	246	0	TRUE	2	0	0	173	130	0	0
59	female	0	2	204	0	FALSE	0	0	0.8	162	134	0	1
58	Male	0	0	283	0	TRUE	2	0	1	162	150	0	0

Table 1: Cleveland Data Set Sample[13]

The Attributes of Cleveland Dataset are considered as

1.Age:Age plays major role in heart disease prediction. As age increases chance of getting heart diseases also increases with every five years of age because of that age considered is in between the range of 29-69.

2.Gender:Gender also plays a vital role in prediction, mens are having more chance of getting the heart diseases.but recent studies shows that females with diabetic also has more chances of getting heart diseases compare with men with same issues

3.Chest pain: Chest pain is also called as angina. Type of chest pain which was caused due to less oxygen in blood supply leads to heart diseases. Values given for different types of chest pain is given below.

1 is for typical angina
2 is for atypical angina
3 is for non- anginal pain
4 is for asymptotic

4.Ca:Number of major vessels coloured by fluoroscopy values given for this attribute ranges from 0 to 3.
5.cholestrol:Two types of cholesterols are there low-density lipoprotein (LDL) and high-density lipoprotein (HDL).High level of LDL leads towards the heart disease so it is better to maintain good value of HDL and less value of LDL. To know this level blood test is needed.unit for cholesterol is Mg/dl

6. Exang: Exercise induced angina or not can be checked. If angina is there after exercise value 1 is assigned else value 0
2 reversable defect
is assigned. angina is the pain occur in the middle of the chest. sometimes it may occur in the shoulders, hands, jaw and neck.

7. Fasting Blood Sugar: Unit for fasting blood sugar is mg/dl if it is morethan 120 then fasting blood sugar is considered as true else it is considered as false.

8. Restecg: This represents electro-cordio-graphic values at rest.

Values are ranged from 0 -2 where 0 indicates normal,1 indicates having ST-T wave abnormality and

2 indicates left ventricular hyperthrophy

9. Slope:slope of the peak exercise ST segment this is nothing but ECG stress test that can be conducted on treadmill.range of values given for this attribute is 1-3 where 1 represents upsloping,2 represents flat and
3 represents downsloping .based on these values an individual can be referred for immediate angiography if ST segment value is morethan 1.

10. oldpeak:ST depression induced by exercise relative to rest
11. Thalach: thalach is Max heart rate achieved by a person. Max heart rate always lead towards the heart diseases, it increases the risk of cardiovascular disease. high blood pleasure and stress is the major reasons behind this.
12. Trestbps: Blood pressure at rest
13. thalassemia: it gives result as normal or not.
0 for normal
1 for fixed defect

S.no	**Title**	**Author**	**Methodology Used**	**Advantages**	**Existing system gaps/Limitations**
1	Data Science And Its Application In Heart Disease Prediction	M. J. A. Junaid et al	Naive Bayes,Artificial neural networks and SVM	88% Acuracy	Less specificity
2	Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques	S. Mohan et al	Random forest with a linear model	Good precision and sensitivity. With 88.7% accuracy	Combination of other model will incresse the accuracy.
3	Data Science Framework - Heart Disease Predictions, Variant Models and Visualizations	C. S. Prakash et al	SVM and logistic regression	Good precision with logistic regression	Only 4 attributes age,sex,chest pain and fasting blood sugar used
4	On Machine Learning Models for Heart Disease Diagnosis	C. -H. Lin et al	CNN	categorical and noncategorical data used.Accuracy 78% with three hidden layers.	If neurons>100 then less accuracy Like 77%
5	HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System	N. L. Fitriyani et a	SMOTE and ENN Synthetic Minority Over-sampling Technique-Edited Nearest Neighbor (SMOTE-ENN) t Synthetic Minority Over-sampling Technique-Edited Nearest Neighbor (SMOTE-ENN) t	They used strong pre-processing model including feature selection ,outlier detection and removal, clustering through DBSCAN. XGBoost Accuracy 98%	Comparision and analysis of different outlier methods was not applied.
6	Heart Disease Prediction using WEKA tool and 10-Fold cross-validation	Marjia Sultana et al	j48,SMO,k-star,bayes net and multilayer perceptron using WEKA tool	SMO got good accuracy of 89%	it is not enough to predict heart disease perfectly With only SMO further improvement of accuracy is needed.
7	Predictive analytics to prevent and control chronic disease	Dr.S.SeemaShedole et al	Naive bayes,SVM and decision tree	SVM provided Accuracy of 94% for heart disease	Not suitable for large dataset only works efficiently with Cleveland dataset
8	Evaluate the performance of different machine learning techniques for prediction of heart disease using ten-fold cross-validation	Ashok kumar Dwivedi et al	Naive bayes, logistic regression ,KNN and classification Tree	Only logistic regression gives more accuracy of 85%	Other methods accuracy is low

9	Heart Disease Prediction System using Data Mining Techniques	Megha Shahi et al	SVM, Naïve Bayes, Association rule, KNN, ANN, and Decision Tree	SVM provides high accuracy among other methods	Not suitable for large dataset only works efficiently with Cleveland Dataset
10	A conceptual method to enhance the prediction of heart diseases using the data techniques	R. Sharmila et al	SVM in parallel fashion	SVM provides better and efficient accuracy of 85% and 82.35%. SVM in parallel fashion gives better accuracy than sequential SVM.	Not suitable for large dataset only works efficiently with Cleveland dataset and also not efficient if missing values are more in database.
11	Different Data Mining Approaches for Predicting Heart Disease	Ashwini Shetty A et al	Neural network and Hybrid systems with datamining techniques	Accuracy of Neural Network 84% and Accuracy of Hybrid Systems 89%	Usage of matlab tool is not that efficient

Table 2: Comparison of Different Studies

This report highlights key areas of research in heart disease prediction, focusing on recent advancements and potential future directions. By exploring these areas, researchers and healthcare professionals can gain valuable insights into emerging techniques and technologies that can enhance early detection, risk assessment, and personalized treatment strategies for heart diseases.

3. CONCLUSION AND FUTURE ENHANCEMENT

Once the heart disease is predicted there are plenty of treatment opportunities to save the life of the person. Using optimum technique for the prediction is important. Here we have provided many techniques and algorithms from machine learning and datamining to predict the heart disease. Use the relevant Pre processing and feature selection technique to get more accurate prediction.

By this survey we conclude that whenever the combination of two or more techniques are used then we get the more accurate prediction, so that hybrid techniques or complex algorithms like CNN or RNN can be optimized because they can extract intricate patterns and relationships from complex medical data, if possible inclusion of more entries are needed to get more accurate results.

Wearable devices, such as smartwatches and fitness trackers, enables continuous monitoring of physiological parameters relevant to heart disease prediction. These devices can collect data on heart rate, blood pressure, physical activity, and sleep patterns, providing valuable insights into an individual's cardiovascular health. Research in this area focuses on developing algorithms and models to analyse wearable data for early detection of abnormalities and real-time risk assessment.

Combining clinical parameters with genetic information can enhance heart disease prediction models. Genetic factors play a significant role in cardiovascular diseases.

There is a scope of improvement of the accuracy and precision if we have used multiple classifiers and different types of decision trees. we can also enhance the existing study by using association rules and logistic regressions. In future we can also use real world dataset instead of the existing dataset. In future work researcher can predict the heart disease and deploy a model for classification later detect the heart disease accurately with deep learning techniques. The combination of feature selection algorithms used for prediction and heart disease severity classification. Continued research in these areas will lead to improved diagnostic accuracy, better management strategies, and ultimately, enhanced patient outcomes in the field of cardiovascular medicine.

REFERENCES

[1]https://healthywa.wa.gov.au/Articles/A_E/Common-medical-tests-to-diagnose-heart-conditions.

[2]M. J. A. Junaid and R. Kumar, "Data Science And Its Application In Heart Disease Prediction," 2020 International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2020, pp. 396-400, doi: 0.1109/ICIEM48762.2020.9160056.

[3] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542- 81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[4] C. S. Prakash, M. Madhu Bala and A. Rudra, "Data Science Framework - Heart Disease Predictions, Variant Models and Visualizations," 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), Gunupur, India, 2020, pp. 1-4, doi: 10.1109/ICCSEA49143.2020.9132920.

[5] C. -H. Lin, P. -K. Yang, Y. -C. Lin and P. -K. Fu, "On Machine Learning Models for Heart Disease Diagnosis," 2020 IEEE 2nd Eurasia Conference on

Biomedical Engineering, Healthcare and Sustainability (ECBIOS), Tainan, Taiwan, 2020, pp. 158-161, doi: 10.1109/ECBIOS50299.2020.9203614.

[6] N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," in IEEE Access, vol. 8, pp. 133034-133050, 2020, doi: 10.1109/ACCESS.2020.3010511.

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

[8] Dr.S.SeemaShedole, Kumari Deepika, “Predictive analytics to prevent and control chronic disease”, https://www.researchgate.net/punlication/316530782, January 2016.

[9] Ashok kumar Dwivedi, “Evaluate the performance of different machine learning techniques for prediction of heart disease using ten-fold cross-validation”, Springer, 17 September 2016.

[10] Megha Shahi, R. Kaur Gurm, “Heart Disease Prediction System using Data Mining Techniques”, Orient J. Computer Science Technology, vol.6 2017, pp.457-466.

[11] R. Sharmila, S. Chellammal, “A conceptual method to enhance the prediction of heart diseases using the data techniques”, International Journal of Computer Science and Engineering, May 2018.

[12] Ashwini Shetty A, Chandra Naik, “Different Data Mining Approaches for Predicting Heart Disease”, International Journal of Innovative in Science Engineering and Technology, Vol.5, May 2016, pp.277-281.

[13] https://www.kaggle.com/ronitf/heart-disease-uci