Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7765978

A COMPREHENSIVE REVIEW OF THE HEART DISEASE PREDICTION CLASSIFICATION USING MACHINE LEARNING

Shatendra Kumar Dubey#1
Ph.D., Research Scholar, RNTU Bhopal, shatendradubey@gmail.com

Dr. Sitesh Sinha#2
Prof. CSE Dept. RNTU Bhopal, MP, India Email: siteshkumarsinha@gmail.com

Dr. Anurag Jain#3
Department of Computer Science & Engineering REC, Bhopal, anurag.akjain@gmail.com

Abstract:
The heart plays an important role in the human body. The diagnosis and prediction of heart-related diseases require greater precision, perfection, and correctness because a minor error can result in exhaustion or death. There are numerous death cases related to the heart, and the number is growing exponentially day by day. This task presents several machine-learning techniques for predicting heart diseases using data on major health figures from patients. The paper shows four classification methods: k-nearest neighbor, support vector machine (SVM), random forest (RF), and linear regression (LR) to build the prediction models. After that, we compare our algorithm with Multilayer Perceptron (MLP), Decision Tree (DT), Random Forest (RF), and Naive Bayes (NB) to build the prediction models. Data pre-processing and feature selection steps were done before building the models. The models were evaluated based on accuracy, precision, recall, and F1 score. The SVM model performed best with 93.17% accuracy. For the implementation, we have used Python programming, which has many types of libraries and header files that make the work more accurate and precise.
Keywords: MLP, SVM, Pre-processing, RF, Linear Regression.

Introduction

One of the main diseases affecting many middle-aged and elderly people is heart disease, and in many cases, it leads to fatal complications. According to WHO statistics, an estimated 24% of deaths from non-communicable diseases in India are due to heart disease. About 17 million people die from cardiovascular disease (CVD) each year worldwide, and the disease is common in Asia. Age, sex, smoking, family history, cholesterol, poor diet, high blood pressure, obesity inactivity, and alcohol intake are considered risk factors for heart disease and other factors. Genetic risks such as high blood pressure and diabetes also lead to heart disease.

It is generally seen that so many times that you or someone closely need doctor help urgently basis, but they are not available sometimes. The Heart Disease Expert System (HDES) is an end-user application for online consultation. Here, we propose a smart system application that allows users to get instant counseling on their heart disease through an online intelligent system. Because of this, we have proper data preparation, and a proper feature selection procedure is required to achieve high accuracy in predicting heart disease using significant features and data mining techniques. Although feature selection is as important as choosing the right technique, researchers still struggle to combine the right data-mining technique with the right set of features. Accurate diagnosis of cardiovascular disease is expected, but it is not an easy task. Moreover, the combination of important features will improve the accuracy of the forecast. This indicates that extensive experimentation to identify key features is required to achieve this goal. Proper evaluation and comparison to test different feature combinations and data mining techniques have not yet been the focus. Therefore, exhaustive experimentation will be required to accurately identify data mining techniques and key features to ensure that heart disease predictions are acceptable and accurate. The application is fed with various details and the heart disease associated with those details. The intelligent smart system allows user to share their heart-related issues. It then processes user-specific details to examine for various illnesses that could be related to it. Here we use a few intelligent data mining techniques to predict the most accurate illness that could be related to a patient's details. Depending on the result, they can contact the doctor for further treatment. The system permits users to view doctor details too. The system can be used for heart disease consulting online. Machine learning (ML) is causing quite a buzz in the healthcare industry as a whole. Payers to healthcare companies around the world are taking advantage of ML today. In this paper, I will demonstrate a use case and show how we can harness the power of ML and apply it to real-world problems.

Literature Survey

There are numerous works have been done related to disease prediction systems using different data mining techniques and machine learning algorithms in medical centers.

Farzana Tasnim et al, This investigation makes use of several data mining classification algorithms to determine the prevalence of cardiovascular diseases, including Naive Bayes, SVM, KNN, DT, NN, LR, RF, and Gradient Tree Boosting. Using a dataset, the estimation of cardiovascular disease is assessed. The success of ML algorithms will be improved by feature selection approaches. With a 92.85% accuracy level, the RF algorithm. For cardiovascular disease, the RF algorithm with PCA performs better than all other algorithms utilized in the current study.

Srinivas et al, Here, we provide a summary of the applications of classification-based data mining methods, including rule-based, DT, NB, and ANN (Artificial Neural Network). Huge amounts of data relating to healthcare are gathered in the sector, but they are not mined to find hidden information. To prepare data and make wise decisions The naïve credal classifier 2 (NCC2) and one dependency-enhanced naive Bayes (ODANB) classifier are employed. The goal of this expanded version of naive Bayes to ambiguous probabilities is to produce reliable classifications even when working with small or insufficient amounts of data. Hidden linkages and patterns are frequently ignored. Heart disease risk can be predicted using a patient's medical history, including blood pressure, diabetes, age, and sex. To identify patterns and links between medical parameters, it identifies and confirms relevant information concerning heart disease. We provide an overview of machine learning methods and their difficulties.

Abhay Agrahara et al, This research report seeks to assess the findings of research on heart disease prognosis. The heart disease prediction system improves in a range of settings by merging Decision Trees and Artificial Neural Networks. This paper provides an overview of widely used data mining and machine learning approaches.

Devansh Shah et al, This article discusses some aspects of heart disease and models based on supervised learning methods such as Nave Bayes, decision trees, K-nearest neighbors, and random forest algorithms. Cleveland Database The UCI Repository of Patients with Heart. The data set has 303 instances, each with 76 attributes. Only 14 attributes are tested to demonstrate how the various algorithms work. The purpose of this study is to assess a patient's risk of developing heart disease. When the results are displayed, the nearest neighbor has the highest accuracy score.

C B Gokulnath et al, The purpose of this paper is to propose an optimization function based on support vector machines (SVM). A genetic algorithm (GA) uses this target function to select features that are more important for the development of heart disease. The results of the GA experimental SVM can be analyzed by a bump, CFS, filtered subset, information gain, coherent subset, chi-square, attribute-based, filtered attribute, gain ratio, and other GA receiver operating characteristics. Contrast with results from feature selection algorithms. An analysis is performed to verify the superior performance of the Evaluated SVM classifier.

C B C Latha et al, This work focuses on improving the robustness of weak classification algorithms and demonstrating their usefulness in prediction. The main goal of this study was not only to improve the accuracy of weak classification algorithms but also to demonstrate how valuable they are in predicting early disease. This study demonstrates ensemble strategies such as increasing the prediction accuracy of weak classifiers. possibility of developing heart disease.

M S Amin et al, The purpose of this study is to find important characteristics and data mining methods that can improve the accuracy of cardiovascular disease prediction. Other researchers used different feature sets and his seven classification methods to create predictive models. It supports Vector Machines (SVM), Decision Trees, Naive Bayes, Logistic Regression (LR), Neural Networks, and Voting (a hybrid technique using Naive Bayes and Logistic Regression). Farman Ali et al, Ensemble deep learning and feature fusion approaches have been proposed for cardiac disease prediction using smart healthcare systems. To create useful health data, the feature fusion approach first merges features collected from sensor data and electronic medical records. Second, information retrieval technology eliminates unnecessary functions and selects important functions to reduce computational load and improve system performance.

Mangesh et al, In this study, we test our approach using the Cleveland heart sample from the UCI library. Our Cluster-Based His DT Learning (CDTL) consists of five main steps. The target labeling distribution was first used to segment the initial set. Other possible class combinations were created from highly distributed samples. Relevant features were determined for each class set combination using entropy. Entropy-based partitions serve an important function. Finally, in these entropy clusters, RF performance is consistent with all important features in predicting heart disease. With our CDTL approach, the HF classifier achieves 89.30% predictive accuracy improvement versus 76.70% accuracy (without CDTL). Therefore, the error rate of HF with CDTL was significantly reduced from 23.30% to 9.70%.

M. Shouman et al, It is difficult to assess the relative accuracy of current data mining techniques in diagnosing cardiac disease from the literature. In this study, we evaluate the performance of Decision Trees, Naive Bayes, and K Nearest Neighbours in diagnosing heart disease patients and further improve their performance by incorporating clustering methods. We performed tests on standardized datasets widely used in the literature to further evaluate the performance improvement of incorporating clustering techniques. Results showed that combining clustering with decision trees, naive Bayes, and k nearest neighbors can improve accuracy. Her ensemble of two clusters in lies k-mean clustering using the k-nearest neighbor method effectively diagnosed cardiac disease.

Divya Tomar et al, In this study, a machine learning technique, Least Squares Twin Support Vector Machine (LSTSVM), is used to diagnose heart disease. Calculate the weight of each feature using the F-score and select the feature based on its weight. We assigned higher weights to features with higher F-scores and used a grid search approach to select the optimal values of the classifier parameters to improve performance. This study used the Cardiac Statistical Protocol disease data set from the UCI repository. We evaluated the performance of the proposed model using different sets of features from different training test datasets. The results show that his LSTSVM model of 11 features achieved the highest accuracy, and the results are very promising compared to other previously proposed approaches.

Md. Khalid Hossain et al, In this article, we explored how to apply various machine learning algorithms to detect heart disease. The research for this article showed a two-step process. First, we converted the heart disease dataset into the format required to run the machine learning algorithm. Collect medical records and other information about a patient from the UCI repository and use the Cardiology dataset to determine if the patient has heart disease. This article presents many valuable results. Confusion matrices validate the accuracy of machine learning algorithms such as logistic regression, support vector machines, K-nearest neighbors, random forests, and gradient-boosted classifiers. Other algorithms are less accurate compared to the Logistic Regression algorithm which is 95% accurate.

Fahd Saleh Alotaibi et al, Improve the accuracy of HF (heart failure) prediction with the UCI Cardiac Dataset. This is done by estimating the likelihood of HF in medical databases using various machine-learning techniques to make sense of the data. According to results and comparative analyses, recent studies have improved the accuracy of predicting heart disease. Combining the machine learning model proposed in this study with medical information systems can predict heart failure and other diseases.

J. Thomas et al, In recent years, heart disease is prevalent and people's lives are at risk. Different people have different blood pressure, cholesterol, and heart rate levels. However, according to medically proven results, blood pressure is 120/90, cholesterol is 72, and heart rate is 72. This article provides an overview of the various classification methods used to predict an individual's level of risk based on age, gender, blood pressure, cholesterol, and pulse rate. Classify patient risk levels using data mining classification techniques such as Naive Bayes, ANN, and Decision Tree.

Himanshu et al, For low variance and high bias, Naive Bayes outperforms ANN with high variance and low bias. ANN suffers from overfitting problems due to low distortion and high variance. This is why ANN performs poorly. Using a low variance and high bias has some advantages as it reduces the time it takes for the training and testing algorithms due to the small data set. However, using small-size data sets also has some drawbacks. Some researchers work on data mining to predict heart disease.

Sanathana Krishna J et al, In this article, we use data mining classification techniques to evaluate medical parameters and classify datasets concerning these parameters to show the percentage probability of developing heart disease. Two major machine learning algorithms, the Decision Trees algorithm, and the Naive Bayes algorithm were used to process datasets in Python programming to show the most accurate algorithms for predicting heart disease.

M. Nikhil Kumar et al, In this article, we use eight algorithms, including Decision Trees, the J48 Algorithm, Logistic Regression, Naive Bayes, Support Vector Machines, Random Forests, Adaboost, and K-Nearest Neighbours, to explore critically important medical fields such as cardiology. play a role. It is one of the leading causes of death worldwide, causing nearly 47% of all deaths. Our goal is to use data mining and machine learning algorithms to perform predictive analytics for heart disease. Explore the different mining and machine learning algorithms in use to determine the best strategy when using additional attributes to improve prediction accuracy.

Table: 1 Comparison of Accuracy with different Techniques.

S.NO	Author	Technique	Accuracy
1	Farzana Tasnim [1]	Naïve Bayes	82
		KNN	89.5
		SVM	82
		Decision Tree	91.59
		Random Forest	92.85
		Linear Regression	81.68
		NN	83
		Gradient Boosting	84
		XGBoost	83.5
2	Abhay Agrahara [3]	KNN	86.09
		SVM	88.04
		Linear Regression	86.09
		Random Forest	96.82
		Decision Tree	98.29
3	C B Gokulnath [5]	SVM	83.7

		Multilayer Perception	78.14
		J48	76.66
		KNN	75.18

Proposed Work

We propose to study all previous systems, as well as various prediction approaches, including (but not limited to)- k-nearest neighbor, Support Vector Machine (SVM), Random Forest (RF), and linear regression (LR) to build the prediction models. We shall use clinical data in our work research. The purpose of the above-stated study is to develop an ensemble approach for prediction. The resulting system would have low false negatives and low overheads.

Table: 2 UCI attributes Data Set

S.N	Attributes	Description	
1	Age	Patients age in years	Continuous
2	Sex	Sex of subject (male- 0, female-1)	Male/Female
3	CP	Chest pain type	Four types
4	Trest bps	Resting blood pressure	Continuous
5	Chol	Serum cholesterol in mg/dl	Continuous
6	FBS	Fasting blood pressure	<or>120mg/dl
7	Rest ecg	Resting Electrocardiograph	Five values

The main objective is to improve the prediction rate by applying a strong classifier. The clinical data set has been used for training and testing. The KSRL collects a large amount of medical data for overloading the process. So, we try to improve the overloading issues in KSRL. To minimize the false positive rate as well as the time consumption rate

DATASET: We have used Cleveland and stat log Cleveland+ Hungarian dataset from UCI [5] machine learning repository.

Pre-processing of Data: We need to clean and remove the missing or noise values from the dataset to obtain accurate and perfect results, known as data cleaning. Using some standard techniques in python 3.8, we can fill in missing and noise values. Then we need to transform our dataset by considering the dataset's normalization, smoothing, generalization, and aggregation. Integration is one of the crucial phases in data pre-processing, and various issues are considered here to integrate. Sometimes the dataset is more complex or difficult to understand. In this case, the dataset needs to be reduced in a required format, which is best to get a good result.

Result Analysis

Analysis of Heart Disease Dataset Figure 1. Target class. Before going to study the performance of considering machine learning algorithms in this research, an analysis of the features of the heart disease dataset will be focused on here. The total attributes is 1025, where not having heart disease is 499 (denoted by 0) and having heart disease 526 (represented by 1), see Figure 1. So, the percentage of not have heart disease is 45.7%, and the percentage of having heart disease is 54.3%, see Figure 1. It is shown that the rate of heart disease is more than the rate of no heart disease.

The accuracy rate is a correct prediction ratio to the total number of given datasets. It can be written as **Accuracy = TP+TN/TP+FP+TN+FN**

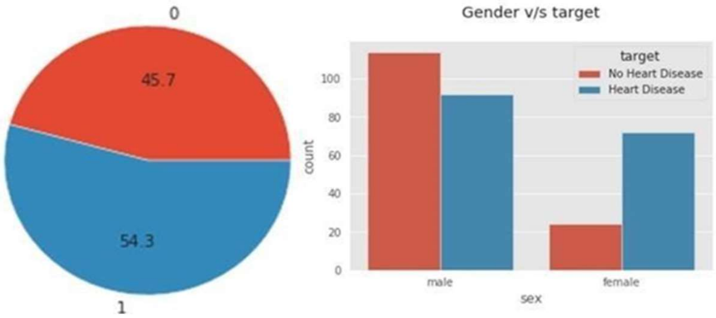

Fig 1 Comparison chart Gender Vs Target

Where TP: True Positive TN: True Negative FP: False Positive FN: False Negative After performing the machine learning algorithms for training and testing the dataset, we can find a better algorithm by considering the accuracy rate.

Experiments and Results

In this section, we discuss the heart disease datasets, experiments, and the evaluation scheme. In this study, we use the Waikato Environment for Knowledge Analysis (Weka)

Classification results

The entire project aimed to test which algorithm classifies heart disease the best with the proposed optimization methods. The classification experiment in this paper was carried out in a Weka Environment. Due to the small number of selected features, 10-fold cross-validation was used. The purpose of avoiding unstable operation results was accomplished by running each experiment 10 times, with the optimal classification accuracy selected for comparison. Some people evaluate the effectiveness of all classifiers in terms of time to build the model, correctly classified instances, incorrectly classified instances, and accuracy according to some steps:

1. Classifiers without optimization
2. Classifiers optimized by FCBF.

Effectiveness

In this section, we evaluate the effectiveness of all classifiers in terms of time to build the model, correctly classified instances, incorrectly classified instances, and accuracy. The results are shown in Table 5 without optimization; Table 5 is optimized by FCBF. To improve the measurement of classifier performance, the simulation error is also taken into account in this study. To do this, we evaluate the effectiveness of our classifier in terms of Kappa as a randomly corrected measure of agreement between classifications and actual classes, Mean Absolute Error as how predictions or predictions approximate possible results, Root Mean Squared Error, Relative Absolute Error, Root Relative Absolute Error, Root Relative Squared Error. The results are presented in Figs. 5, 6.

Table 5. Classifier's Performance without optimization

Evaluation criteria	Kappa Statistic	Mean Absoluteerror	Root mean Squared Error	Relative Absolute Error	Root Relative squarederror
KNN	0.001	0.007	0.016	0.017	0.008
SV M	2.02	2.26	2.2	2.21	2.3
NB	0.68	0.44	0.5	0.7	0.75

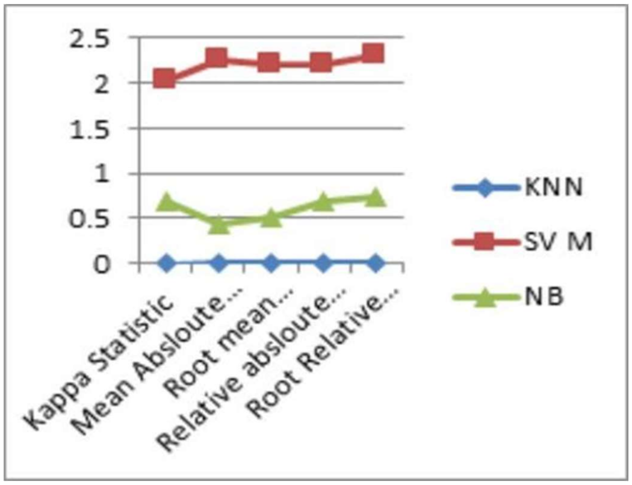

Fig 2 -Classifier's performance without optimization

Conclusion and Future scope

The heart is a vital organ in the human body; however, heart disease is a major concern world because this disease is increasing day by day. So, we can handle this disease if we have a model which can predict the initial condition of heart disease. So, we need to create a machine learning model that can be more accurate and help to diagnose heart disease with less doubt and cost. It can be a primary technique for knowing the condition of the heart. For this reason, this article focuses on heart disease prediction based on the accuracy rate of the confusion matrix. Following this idea, the statistics of the given algorithms are used to estimate the accuracy rate of the confusion matrix and validated the statistics among the machine learning algorithms. When algorithms are compared, it is found that the linear regression algorithm is selected regarding the performance of a high accuracy rate. Data pre-processing and feature selection steps were done before building the models. The models were evaluated based on accuracy, precision, recall, and F1 score. The SVM model performed best with 93.17% accuracy. For the implementation, we have used WEKA TOOL.

References

1. Farzana Tasnim, Sultana Habiba 2021 A Comparative Study on Heart Disease Prediction Using Data Mining Techniques and Feature Selection DOI:10.1109/ICREST51555.2021.9331158.
2. Srinivas, K., Rani, B.K., Govrdhan, A., 2010a. Applications of data mining techniques in healthcare and prediction of heart attacks. Int. J. Comput. Sci. Eng. (IJCSE) 2 (02),250–255.
3. Abhay Agrahara 2020 Heart Disease Prediction Using Machine Learning Algorithms. International Journal of Research in Computer Science, Engineering and Information Technology. ISSN 2456-3307 page137-149.
4. Devansh Shah, Samir Patel, Santosh Kumar Bharti “Heart Disease Prediction using Machine Learning Techniques” SN Computer Science (2020) 345 https://doi.org/10.1007/s42979-020- 00365-y.
5. C B Gokulnath, S.P. Shantarajah “An optimized feature selection based on genetic approach and support vector machine for heart disease “Cluster Computing (2019)22: S14777–S14787.
6. C.B.C. Latha, S.C. Jeeva, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, Informatics Med. Unlocked 16 (2019)100203.
7. Mohammad Shafenoor Amin, Yin Kia Chiam, Kasturi Dewi Varathan “Identification of significant features and data mining techniques in predicting heart disease” Telematics and Informatics 36 (2019) 82-93.
8. Farman Ali, Shaker El-Sappagh, S.M. Riazul Islam, Daehan Kwak, Amjad Ali, Muhammad Imran, Kyung-Sup Kwak “A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion” Information fusion63 (2020)208-222.
9. G. Magesh, P. Swarnalatha “Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction” Evolutionary Intelligence (2021) 14:583–593 https://doi.org/10.1007/s12065-019-00336-0
10. Shouman M., Turner T., Stocker R., 2013. Integrating clustering with different data mining techniques in the diagnosis of heart disease. J. Comput. Sci. Eng. 20 (1). fusion63 (2020)208- 222
11. Divya Tomar and Sonali Agarwal “Feature Selection based Least Square Twin Support Vector Machine for Diagnosis of Heart Disease” International Journal of Bio-Science and Bio-Technology Vol.6, No.2
12. Md. Khalid Hossein “Heart Disease prediction Using Machine learning Techniques” AJCSTISSN:2640-0111 (Print); ISSN:2640-012X(Online)
13. Fahd Saleh Alotaibi,” Implementation of Machine Learning Model to Predict Heart disease”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.

14. Theresa Princy R, J. Thomas, 'Human Heart Disease Prediction System using Data Mining Techniques, International Conference on Circuit Power and Computing Technologies, Bangalore, 2016.

15. Himanshu Sharma and M A Rizvi, "Prediction of Heart Disease using Machine Learning Algorithms: A Survey" International Journal on Recent and Innovation Trends in Computing and Communication Volume: 5 Issue: 8, IJRITCC August 2017.

16. Santhana Krishnan J and Geetha S, "Prediction of Heart Disease using Machine Learning Algorithms" ICIICT, 2019.

17. M. Nikhil Kumar, K. V. S. Koushik, K. Deepak, "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools" International Journal of Scientific Research in Computer Science, Engineering, and Information Technology, IJSRCSEIT 2019.

18. Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronary heart disease prediction model: the Korean Heart Study. BMJ open, 4(5), e00502.