Features Contributing Towards Heart Disease Prediction Using Machine Learning

Chetan Sharma[a], Shankar Shambhu[b], Prasenjit Das[b], Shaily Jain[c], Sakshi[d]

[a]*Chitkara University Himachal Pradesh, India*
[b]*Chitkara University School of Computer Applications, Chitkara University, Himachal Pradesh, India*
[c]*Chitkara University Institute of Engineering and Technology, Chitkara University, Himachal Pradesh, India*
[d]*Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract

WHO and other health organizations claimed that the death rate due to cardiovascular disease is one-third of worldwide. Although, many researchers have worked in this direction to help our medical professionals diagnose this disease at an early stage. This paper aims to apply data mining algorithms to predict heart disease occurrence in patients based on some features like diabetes, blood pressure, etc. We have implemented two data mining algorithms, Naive Bayes and NB tree, on two data different datasets of the UCI repository to evaluate the accuracy, f-measure, precision, and recall. Our results show NB tree outperforms with 84.6% accuracy compared to Naive Bayes with only 80.58 % accuracy.

Keywords: Machine Learning, Classification, Heart, Disease, WEKA

1. Introduction

The heart is the essential central part of the human body, which provides the purified blood to each part of the body. Without a healthy working heart, a person cannot live a single second. But, nowadays, heart diseases are increasing at a rapid speed. As per the WHO, over 17.9 million people died every year because of heart disease, and 80% of people died because of a heart attack [1]. Heart disease has been recognized as one of the world's most complex and life-threatening human diseases. Typically, the heart is unable to push the necessary amount of blood to other areas of the body to satisfy the body's normal functioning. Because of this, heart failure eventually occurs[2]. In the United States, the incidence of heart illness is very high [3]. Swelling in the feet, Chest pain, breathe shortness, body tiredness, Pain in the neck and shoulders, etc., are some significant symptoms of heart disease [4]. Techniques used to diagnose heart diseases at an early stage have been complicated, and the resulting difficulty is one of the critical factors affecting the standard of living [5]. Because of the low availability of instruments and lack of physician, diagnosis of heart diseases and their treatment is very involved in developing countries [6]. It affects the prediction results and treatment of heart

ACI'21: Workshop on Advances in Computational Intelligence at ISIC 2021, February 25-27, 2021, Delhi, India
EMAIL: chetan.sharma@chitkarauniversity.edu.in (C. Sharma); shankar.shambhu@chitkarauniversity.edu.in (S. Shambhu); prasenjit.das@chitkarauniversity.edu.in (P. Das); shaily.jain@chitkarauniversity.edu.in (S. Jain); sakshi@chitkara.edu.in (Sakshi)
ORCID: 0000-0001-5401-8503 (C. Sharma); 0000-0002-2348-1041(S. Shambhu); 0000-0002-7988-2418 (P. Das); 0000-0001-6078-3607 (S. Jain); 0000-0002-8757-4001 (Sakshi)

patients, which is the main reason for the high mortality rate of heart patients. Hence, to reduce the mortality rate of heart patients and provide the best treatment of heart diseases, appropriate and accurate heart disease diagnosis techniques are required [7]. These techniques should be capable of detecting heart disease at an early stage [19-20]. The rest of the paper is organized as follows: Section 2 discusses the background and history of this work, Methodology is explained in section 3 along with the description of tools, datasets, and algorithms used in evaluation, evaluation matrices etc, Results are discussed in section 4 and finally section 5 gives us conclusion of the research done in this paper.

2. Literature Review and Related Work

In the last decade, many researchers worked on heart disease datasets to predict heart diseases. They used multiple machine learning and data mining algorithms for the implementation and achieved different results. Yet today, we also face a lot of issues with heart disease. Following are the literate review of recent research:

The authors implemented three different algorithms Naive Bayes(NB), Artificial Neural Network, and J48 to find the best heart disease prediction results. Researchers used a dataset of 8 additional attributes and 210 instances of male persons. WEKA tool was used for the implementations of the algorithms. Archived results have shown that the Naive Bayes algorithm provided the best results compared to Artificial Neural Network and J48. Naive Bayes achieved an accuracy of 79.90% and took 0.01 second to build the model, where J48 attained the accuracy of 77.03% and took 0.01 second to build the model. Artificial Neural Network achieved an accuracy of 76.55% and took 1.55 seconds to build the model [8].

Kodati et al. used Naive Bayes, Random Forest, J48, and Decision table classification techniques on a dataset of 14 attributes and 303 instances. Used dataset for implementations is taken from UCI repository, and researchers use WEKA tool for implementation. From the implementation results, it has been concluded that the Decision table achieved the best accuracy of 84.81% as compared to the other three algorithms. Naive Bayes(83.70%), Random Forest(81.85%) and J48 achieved an accuracy of 76.66% [9].

Singh et al. developed a new hybrid model named "Hybrid Genetic Naive Bayes Model". This model was developed with two different supervised techniques (Naive Bayes, Genetic Algorithm) for the correct prediction of heart diseases. To develop this model, the researcher used a dataset taken from the UCI repository with 303 instances and 14 important attributes. Implementation results gave the accuracy of 97.14% with 98% precision value and 97.14% recall value [10].

Krishnan et al. used two machine learning algorithms, Decision Tree and Naive Bayes algorithms, to predict Heart Diseases. They used a dataset of 300 instances and 14 attributes taken from the UCI repository. Researchers implemented the python programming language model and achieved the highest accuracy of 91% with a Decision tree and 87% with Naive Bayes [11].

3. Methodology

3.1 Proposed Work

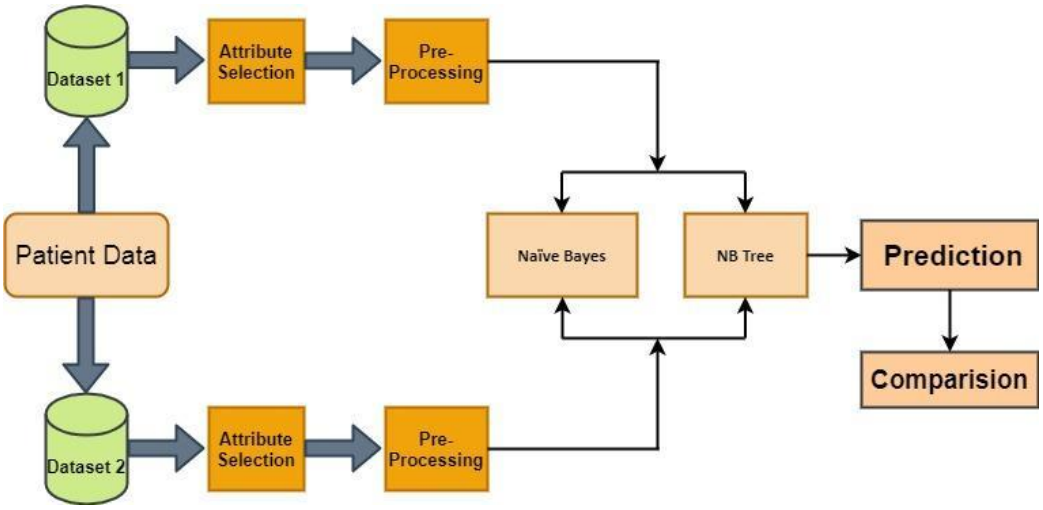

Figure 1: Proposed Methodology of Study

3.2 Tool Used

WEKA 3.8.4 machine learning tool is used to conduct this study, written in Java and developed at the University of Waikato. WEKA tool provides us with different classifiers to examine the performance. WEKA is used to evaluate other data mining tasks like preprocessing, classification, regression, and many more. WEKA accepts .csv and .arff file format and the chosen dataset has already created the required data in the mentioned format.

3.3 Data Preprocessing

The real-life data consists of redundant values and lots of noise. The data needs to be cleaned, and the missing values need to be filled before the data is fed to generate a model [18]. In the preprocessing process, these issues are taken care of so that the prediction can be made accurately. Once the cleaning of data is done, i.e., the noise is removed, and the missing values are filled, we need to transform it. Many supervised learning algorithms work on nominal or cardinal data. So data transformation is applied to the dataset obtained from UCI in the present work. Reduction of the dataset is applied to convert the complex dataset into a more straightforward form, which improves the accuracy of the model.

3.4 Classification Algorithms

After going through an intensive literature review, we have selected two classification algorithms: naive Bayes tree, naive Bayes classification based on their dependency on attributes.

Naive Bayes Tree [12]: It is a hybrid approach in which the model is generated using the Naive Bayes and Decision tree Approach. The naive Bayes classification assumes that the features are independent of each other, and the decision tree assumes that the components are dependent on each other. So the hybrid approach takes advantage of both approaches. The decision tree is built by considering only one feature, and output is fed to the node. Based on the outcome of each node, other features are selected. In this hybrid approach, the split is done in the same manner by considering only one feature at every node but with Naive-Bayes classifiers at the

leaves. In large datasets, data splitting is regarded as a vital and essential task for classification using the features we have implemented the naive Bayes tree classification.

Naive Bayes Classification [13]–[15] : This classification technique is based on Baye's theorem, which works on the assumption that the existence of one feature is independent of the other feature. The advantage of the Naive Bayes classification is that it requires a small amount of data to create/train the model.
Bayes theorem provides a way of calculating posterior probability (conditional probability where we are finding probability under a given condition assumed to be confirmed) P(c|x) from P(c), P(x), and P(x|c). The following is the formula to calculate posterior probability:

$$P(c|x)=P(x|c)*P(c)/P(x|c)$$

Where:
P(c|x) is the conditional probability that occurs when x has already occurred
P(c) is the known probability of the class.
P(x|c) is the conditional probability of x condition that c has occurred.
P(x) the known probability of the class.

Two datasets were used in this study. The first one was obtained from the "Cleveland Clinic Foundation", the First dataset comprises 303 instances. The second dataset is taken from the public available platform, a combination of five other datasets named Heart Disease Dataset (Comprehensive). All the dataset are available for heart disease having a total of 76 attributes and each dataset choose their dataset features accordingly. Initially, both the dataset was selected for the study with 76 attributes, but they were preprocessed to produce 14 and 11 characteristics to reduce redundant variables. Consequently, we used these specific attributes (listed in Table 1 and Table 3) to compare.

The first dataset is taken as the Cleveland database, which is publically available at [16]. There are 303 instances in the dataset, and their description is given in Table 1, and the results using the WEKA tool are given in Table 2.

Dataset Description

Table 1: Cleveland Dataset Attribute Information

Attribute Used	Attribute Information
Age	Age of Patient. The value ranges from 29 years to 77 years
Sex	Gender of the patient represented in binary form 1 = male. 0 = female
Chest Pain	Chest pain. Its value range from 1 to 4. 1 used to represent typical angina, 2 used to describe atypical angina, 3 used to represent non-anginal Pain, and 4 is used to represent asymptomatic.
Resting Blood Pressure	The attribute is used to represent the patient's resting BP, and the unit to measure it is mm Hg.
Cholesterol	The attribute is used to represent the patient's serum cholesterol, and its unit of measurement is mg/dl.
Fasting Blood Sugar	An attribute represents the Fasting blood sugar of the patient. There are two values used in the dataset if the recorded value is > 120 mg/dl, then it is shown by 1 (true), else it is shown by 0 (false). 1 = True. 0 = False.
Resting ECG	The attribute is used to represent the resting electro-cardiographic records of the patient. The value ranges from 0 to 2

	0 is representing the Normal range. 1 is representing the ST-T wave abnormality of the patient. 2 is used to show probable or definite left ventricular hypertrophy by Estes' criteria.
Heart Rate	The attribute is used to represent the maximum heart rate of the patient achieved.
Exercise Included Angina	Exercise-induced angina and represented in binary 1 is used to represent yes. 0 is used to represent no.
Old Peak	The attribute is used to represent ST depression induced by exercise, which is relative to rest.
Slope	The attribute is used to measure the slope for peak exercise. The range of the recorded values is from 1 to 3. Up sloping is represented by 1, flat is shown through value 2, and 3 is used to represent downsloping.
Major Vessels	The attribute is used to represent the no. of significant vessels colored by fluoroscopy. Recorded values are range from 0 to 3, and the value is related to the darkness of the color.
Thallium Scan	The attribute is used to record the Thallium Scan of the patient. It represents the values 3, 6, or 7. 3 represents a normal range, 6 is used to represent fixed defect, and 7 represents reversible defect.

Table 2: Cleveland Dataset Results

Algorithms	Accuracy (%)	F-Measure (%)	Precision (%)	Recall (%)	Time (In Seconds)
NB Tree	84.46	84.5	84.5	84.5	0
Naive Bayes	80.58	80.6	80.6	80.6	1.57

The second dataset is taken from [17], collected from five other heart disease databases. There is a total of 1190 instances in the dataset, and these instances are collected from the dataset Cleveland heart disease dataset instances taken 303, Hungarian heart disease dataset instances have taken 294, Switzerland heart disease dataset instances have taken 123, Long Beach VA heart disease dataset instances have taken 200 and Stalog heart disease dataset instances taken 270. Dataset is a combination of 11 common features between all the datasets. Description of all feature used in the dataset is given in Table 3, and their results using the WEKA tool is given in Table 4.

Table 3: Heart Disease Dataset (Comprehensive) Attribute Information

Attribute Used	Attribute Information
Age	Age of Patient. The value ranges from 28 years to 77 years
Sex	Gender of the patient represented in binary form 1 = male. 0 = female
Chest Pain	Chest pain. Its value range from 1 to 4. 1 used to represent typical angina, 2 used to represent atypical angina, 3 used to represent non-anginal Pain, and 4 is used to represent asymptomatic.
Resting BP	The attribute is used to represent the patient's resting BP, and the unit to measure it is mm Hg.
Cholesterol	The attribute is used to represent the patient's serum cholesterol, and its unit of measurement is mg/dl.
Fasting Blood Sugar	An attribute represents the Fasting blood sugar of the patient. There are two values used in the dataset if the recorded value is > 120 mg/dl then it is shown by 1 (true), else it is shown by 0 (false). 1 = True. 0 = False.
Resting ECG	The attribute is used to represent the resting electro-cardiographic records of the patient. The value ranges from 0 to 2 0 is representing the Normal range. 1 is representing the ST-T wave abnormality of the patient. 2 is used to show probable or definite left ventricular hypertrophy by Estes' criteria.
Maximum Heart Rate	The attribute is used to represent the maximum heart rate of the patient achieved.
Exercise Angina	Exercise-induced angina and represented in binary 1 is used to represent yes. 0 is used to represent no.
Old Peak	The attribute is used to represent ST depression induced by exercise, which is relative to rest.

ST Slope	The attribute is used to measure the slope for peak exercise. The range of the recorded values is from 1 to 3. Up sloping is represented by 1, flat is shown through value 2, and 3 is used to represent downsloping.
Target	Used for the prediction

Table 4: Heart Disease Dataset (Comprehensive) Results

Algorithms	Accuracy (%)	F-Measure (%)	Precision (%)	Recall (%)	Time (In Seconds)
NB Tree	88.39	88.4	88.4	88.4	5.54
Naive Bayes	83.70	83.7	83.7	83.7	0

3.5 Evaluation Matrices

We have considered four parameters for our paper. In the present work, the prediction class is if the person having specific attributes has died because of heart disease or not, so the class C in the above table is no. of instances belonging to the class. Figure 2 is the confusion matrix.

TP is the actual no of people who died because of heart disease, and the model also predicted the same. Similarly, TN is the person who didn't die of a heart ailment, and our model also predicted the same. False Positive (FP) is a Type I error because the model predicted that the person died of an ailment, but actually, the patient didn't. False-negative is a type II error. The model predicted that the person didn't die of the alignment, but actually, he/she did.

The accuracy of the model is calculated through the formula given below:

Accuracy = (TP+TN)/Total no. of instance (1)

Recall is the measure of correctly predicted classes out of the total positive classes. The formula is as follows:

Recall= (TP)/(TP+FN) (2)

Precision is the measure of actual positive classes out of all the correctly predicted positive classes. The formula for the recall is as follows:

Precision = TP/(TP+FP) (3)

Comparing the two models becomes problematic when the precision is low, and the recall value is high. In the case of vice versa is true. The two parameters are not of much use for comparison of the models. F-score is used to compare the models in such cases. F-score uses the harmonic mean of the two values. This helps to measure the recall and precision at the same time. Instead of the Arithmetic mean, harmonic mean is used because Arithmetic mean is sensitive to extreme values.

F-score= (2*Recall*Precision) / (Recall + Precision) (4)

Actual class\Predicted class	C	Not in C
C	**True Positives (TP)**	**False Negatives (FN)**
Not in C	**False Positives (FP)**	**True Negatives (TN)**

Figure2: Confusion Matrix

4. Results and Discussion

We have used two datasets with 303 instances in the present work in the first and 1190 in the second set. Naive Bayes and Naive Bayes tree Algorithm has been applied on the two datasets. We find that the NB tree performs better in the two datasets, which are of different sizes and attributes. The accuracy and other measures are better in the NB tree case, which is a hybrid of Naive Bayes and Decision tree. We have applied these two algorithms because the Naive Bayes Algorithm works on the hypothesis that the features are independent of each other. At the same time, the decision tree assumes that the features are dependent on each other. The present work tries to determine if the parameters age, gender, cholesterol, etc., do contribute towards heart disease, and a machine learning algorithm can be used to predict the alignment based on these parameters with an accuracy of 88%.

5. Conclusion

The two datasets used in the present work show a similar accuracy, which leads us to conclude that the machine learning algorithms can predict heart diseases in patients with specific existing alignments like High BP, cholesterol, etc. We find a difference in the accuracy of the two methods applied on the two datasets, namely Naive Bayes and NB tree. The difference in accuracy is that Naive Bayes assumes the independence of features. NB Tree (a hybrid of the Decision tree) assumes that the features are dependent on each other. Higher accuracy in the NB tree makes us conclude that parameters like age, gender, cholesterol, and high Bp are dependent on each other, leading to a heart ailment in patients.

References:

[1] W. H. O. (WHO), "Cardiovascular Diseases." https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (accessed Nov. 15, 2020).

[2] A. L. Bui, T. B. Horwich, and G. C. Fonarow, "Epidemiology and risk profile of heart failure," Nat. Rev. Cardiol., vol. 8, no. 1, p. 30, 2011.

[3] P. A. Heidenreich et al., "Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association," Circulation, vol. 123, no. 8, pp. 933–944, 2011.

[4] M. Durairaj and N. Ramasamy, "A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate," Int. J. Control theory Appl., vol. 9, no. 27, 2016.

[5] J. Mourao-Miranda, A. L. W. Bokde, C. Born, H. Hampel, and M. Stetter, "Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data," Neuroimage, vol. 28, no. 4, pp. 980–995, 2005.

[6] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, "Innovative artificial neural networks-based decision support system for heart diseases diagnosis," 2013.

[7] F. Amato, A. López, E. M. Peña-Méndez, P. Va\vnhara, A. Hampl, and J. Havel, “Artificial neural networks in medical diagnosis.” Elsevier, 2013.

[8] S. K.Gomath, "Heart Disease Prediction Using Data Mining Classification," Int. J. Res. Appl. Sci. Eng. Technol., vol. 4, no. 2, 2016, doi: 10.18775/ijmsba.1849-5664-5419.2014.43.1004.

[9] R. V. Sarangam Kodati, "A Comparative Study on Open Source Data Mining Tool for Heart Disease," Int. J. Innov. Adv. Comput. Sci., vol. 7, no. 3, 2018, [Online]. Available:

http://www.diva-portal.org/smash/get/diva2:1080911/FULLTEXT01.pdf.

[10] N. Singh, P. Firozpur, and S. Jindal, "Heart disease prediction system using hybrid technique of data mining algorithms," Int. J. Adv. Res. Ideas Innov. Technol., vol. 4, no. 2, pp. 982–987, 2018.

[11] S. Krishnan and S. Geetha, "Prediction of Heart Disease Using Machine Learning Algorithms.," in 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), 2019, pp. 1–5.

[12] S. Wang, L. Jiang, and C. Li, "Adapting naive Bayes tree for text classification," Knowl. Inf. Syst., vol. 44, no. 1, pp. 77–89, 2015.

[13] L. Li, Y. Wu, and M. Ye, "Experimental comparisons of multi-class classifiers," Informatica, vol. 39, no. 1, 2015.

[14] P. Ahmad, S. Qamar, and S. Q. A. Rizvi, "Techniques of data mining in healthcare: a review," Int. J. Comput. Appl., vol. 120, no. 15, 2015.

[15] S. S. Nikam, "A comparative study of classification techniques in data mining algorithms," Orient. J. Comput. Sci. Technol., vol. 8, no. 1, pp. 13–19, 2015.

[16] Ronit, "Heart Disease UCI," 2018. https://www.kaggle.com/ronitf/heart-disease-uci (accessed Nov. 12, 2020).

[17] M. Siddhartha, "Heart Disease Dataset (Comprehensive)," 2019. https://www.kaggle.com/sid321axn/heart-statlog-cleveland-hungary-final (accessed Nov. 12, 2020).

[18] V. Madaan and A. Goyal, "Predicting Ayurveda-Based Constituent Balancing in Human Body Using Machine Learning Methods," in IEEE Access, vol. 8, pp. 65060-65070, 2020, doi: 10.1109/ACCESS.2020.2985717.

[19] Vishu Madaan and Anjali Goyal, “Analysis and Synthesis of a Human Prakriti Identification System Based on Soft Computing Techniques”, Recent Patents on Computer Science, 12(1), pp 1-10, 2019. DOI: 10.2174/2213275912666190207144831

[20] Prateek Agrawal, Vishu Madaan, Vikas Kumar, “Fuzzy Rule Based Medical Expert System to Identify the Disorders of Eyes, ENT and Liver”, International Journal of Advanced Intelligence Paradigm (IJAIP), vol 7, issue3-4, pp. 352-367, Inderscience Publications, 2015.