ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 113-118

Article Title: Heart Disease Prediction Using Various Classification Models

Heart Disease Prediction Using Various Classification Models

R. Raja Aswathi[1*], K. Pazhani Kumar[2], B. Ramakrishnan[3]

[1*]Department of Computer Science, S.T. Hindu College, Nagercoil, India. rajaaswathi@outlook.com

[2]Head of the Department, Department of Computer Science, S.T. Hindu College, Nagercoil, India. skpk73@gmail.com

[3]Associate Professor, Department of Computer Science, S.T. Hindu College, Nagercoil, India. ramsthc@gmail.com

ABSTRACT

Heart disease can be prevented with accurate prediction, but it can also be fatal if the prediction is erroneous. The results and characterization of the UCI Machine Learning Heart Disease dataset are investigated in this research using various machine learning and deep learning mechanisms. This study includes the ensemble of methods, well-known algorithms, comparisons with other better methodologies, using an efficient feature selection technique, hybrid approach, fuzzy-based algorithms, removing the noisy data using an enhanced approach, and so on. The dataset contains 14 key attributes that were used in the assessment. The precision of machine learning algorithms is determined by the dataset used for training and testing. The knowledge saved can be useful as a source for anticipating future illnesses. The purpose of this study is to summarize fresh research together with relative outcomes on coronary health risk, as well as to encourage innovative goals using data mining and machine learning frameworks.

Keywords: Classification, Decision Tree, C4.5, Data Mining, Neural Network, Distance-Based Mining

1 Introduction

Coronary illness has sparked a lot of genuine concern among researchers. One of the most challenging aspects of coronary disease is accurately recognizing and also identifying its presence inside an individual. Initial practices are no longer as effective as they once were in recognizing it, and clinical trainers are less skilled in anticipating heart illness. There are several clinically available technologies in the display for predicting coronary artery disease, as well as two major concerns underlying them. One is that they are generally expensive, and another is that they are not effectively capable of calculating the risk of coronary sickness in particular. Gender, age, obesity, increased blood pressure, hyperlipidemia, diabetes, family history, heavy drinking, and smoking are all risk factors for heart disease. Apart from these, many additional hazards have increased in our technical and modern society, such as underemployment industrialization, discomfort extreme tiredness, lack of exercise, anxiety, hypertension, eating adjustments, and sleeplessness. Lifestyle choices, self-interest, ambition, and an egocentric attitude among people have all contributed to the emergence of infectious conditions in people [1].

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 113-118

Article Title: Heart Disease Prediction Using Various Classification Models

Scientific technology is one of the areas where the measurement of software technology can be used, as advancement in machine learning has opened up enormous opportunities in numerous fields. Clinical science also made use of some of the most important tools available in personal computer technology. An expert system gained traction a decade ago due to advancements in computation ability. Machine learning is one of the most widely utilized technologies in a variety of fields because it does not require unique calculations for each database [2]. Figure 1 represents some of the methodologies that are employed in the Coronary heart disease dataset.

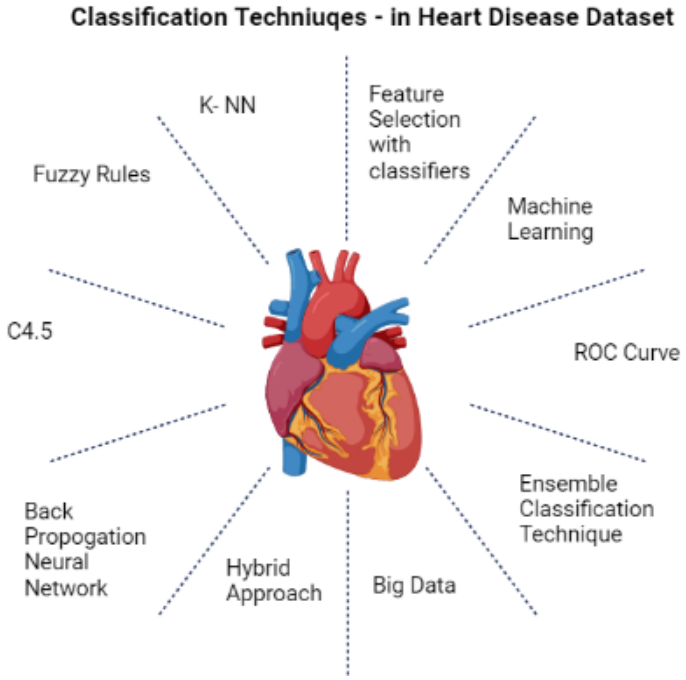

Figure 1: *Employment of heart disease data in different approaches*

The high dimensionality of data is a prevalent challenge in machine learning; the datasets we use contain large amounts of data, and we sometimes can't see it even in 3D, which is known as the curse of dimensionality [3].

As a result, when we execute operations on this data, we need a lot of memory, and the data can grow exponentially, resulting in overfitting. The weighting features can be utilized to reduce redundancy in the dataset, which in turn helps to reduce the processing time of the execution [4][5]. The following survey is carried out with the performances of the heart disease dataset with new terminologies of Artificial Intelligence.

2 Classification Methods

There are various classification methods to predict heart disease using the provided technique, but only some classification algorithms work fine in terms of accuracy and efficiency. Here some methods analysed for good disease prediction at the early stage are discussed.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 113-118

Article Title: Heart Disease Prediction Using Various Classification Models

2.1 Ensemble Based on Distances for a KNN

A.P. Pawlovsky [6] developed an ensemble-based distance for kNN (k-Nearest Neighbor) method is implemented. The ensembles evaluate the average accuracy of the kNN method with the Euclid, Manhattan, Chebyshev, Sorensen, Canberra, and Mahalanobis distances; one uses three distances and another uses five with a weighted version to produce an accuracy of 84.83%. The accumulation of various distance metrics causes concern because not all distance methods are suited well for a single type of data. Here the Mahalanobis performs well than other metrics; however, the solo performance of other distance-based metrics is low. So when collided with other metrics it reduces the overall result.

2.2 Application of Machine Learning

P. S. Kohli et al. [7] included data munging and attributes selection process for datasets. The machine learning algorithms including Logistic Regression, Decision Trees, Random Forest, Support Vector Machine (SVM), and Adaptive Boosting, are used for the prediction of diseases. A comparison of accuracy is done for selecting the best model for the disease dataset. The prediction accuracy is 87.1% in Heart Disease detection using Logistic Regression. Using a pipeline structure for data pre-processing could further help in improved results.

2.3 Fisher score and Matthews correlation coefficient-based feature subset selection

S. M. Saqlain et al. [8] three algorithms are proposed for selecting candidate feature subsets: (1) mean Fisher score-based feature selection algorithm, (2) forward feature selection algorithm and (3) reverse feature selection algorithm. The K-fold validations confirm the results achieved by the proposed technique are the most accurate with an accuracy of 82.7%. Not only with K-fold validation but also with accuracy, sensitivity and specificity can improve further to stand out among other featured algorithms.

2.4 Machine Learning Algorithms with ROC Curve

R. Kannan et al. [9] process uses four distinct machine learning classification techniques, like logistic regression, stochastic gradient boosting, Random Forest, and support vector machines are compared. The accuracy of machine learning algorithms for predicting and diagnosing is done via a receiver operating characteristic (ROC) curve. The greatest performer is logistic regression, which can predict with an accuracy of 87% percent. However, a better-extended version of the base logistic algorithm can efficiently increase the accuracy and speed.

2.5 Ensemble classification techniques

C. B. C. Latha et al. [10] compares classification techniques such as decision tree's C4.5, Random Forest (RF), Naive Bayes (NB), Bayes Network (BN), and Neural Networks. The Supervised Neural Network-based Multiple Perceptron (MP) can be used for the diagnosis of heart disease. Ensemble approaches have been used to increase classification accuracy in prediction. The ensemble algorithms bagging, boosting, stacking, feature selection, and majority voting were used for the experiments. The proposed model has a majority vote with

NB, BN, RF, and MP 85.48%. As the outcome is favorable in an efficient way, the addition of the proposed ensemble methods increases the memory consumption and speed of execution.

2.6 Feature Selection Approaches

S. Bashir at al. [11] the Rapid miner as a tool and Minimum Redundancy Maximum Relevance Feature Selection (MRMR); Decision Tree, Logistic Regression, Logistic Regression (SVM), Naïve Bayes and Random Forest algorithms are extended and examined to find the better one. Logistic Regression is the best feature selection technique with 84.85% accuracy. Logistic Regression can be used as an ensemble with multiple techniques and be worked with real-time datasets for higher performance.

2.7 Using SVM Machine Learning Algorithm

N. Louridi et al. [12] improve the quality by using a better pre-processing phase. It is compared with different Learning algorithms using accuracy, precision, f1-score, and recall performance metrics. The accuracy of 86.8% was obtained by using SVM with a linear kernel. Even including a new approach can improve this domain.

2.8 Classification for Computer Diagnosis System

K. Sathya et al. [13] using WEKA 3.8 software, examines multiple classifier methods such as SVM, KNN, and MLP for a dataset of seer heart disease. Classifier performance is measured using metrics such as classification accuracy, precision, recall, F-Measure, ROC, and Matthews correlation coefficient (MCC). With an accuracy of 85.9%, SVM is the most accurate classification algorithm. A stronger model with an ensemble would permit a far better outcome.

2.9 Using kNN Machine Learning Algorithm

A. Singh et al. [14] proposed machine learning methods for predicting cardiac disease including a k-nearest neighbor, decision tree, linear regression, and support vector machine (SVM). Anaconda (jupytor) notebook is the finest tool for implementing Python programming since it has various types of libraries and header files that make the task more exact. Based on the confusion matrix, it is determined that KNN is the best option with an 87%. More machine learning approaches can be employed for the best analysis of heart diseases.

3 Conclusion

The study covers the recent methodologies on heart disease in classification using Machine Learning classification algorithms, feature selection, ROC Curve, Ensemble classification, big data, hybrid approaches, Neural Network algorithms, C4.5, Fuzzy Rules, etc. Each methodologies found in the review part contains various advantages in research field of classification and medical science. From the analysed techniques new features additions and relationship of work done are understood to increase the information gain in the literature survey. In future, the best performing methodology can be chosen and be upgraded for better efficient outcome in predicting heart disease.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 113-118

Article Title: Heart Disease Prediction Using Various Classification Models

References

1. Jagmohan Kaur; Baljit S. Khehra, Year: 2021, "Fuzzy Logic and Hybrid based Approaches for the Risk of Heart Disease Detection: State-of-the-Art Review", J. Inst. Eng. India Ser. B.

2. Simran Verma; Abhishek Gupta, Year: 2021, "Effective Prediction of Heart Disease Using Data Mining and Machine Learning: A Review", in: 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS). Presented at the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), IEEE, pp. 249 – 253.

3. Claude Sammut; Geoffrey I. Webb, Eds, Year: 2010, "Encyclopedia of Machine Learning", Springer US, Boston, MA.

4. Maryam Imani; Hassan Ghassemian, Year: 2015, "Feature Extraction Using Weighted Training Samples", IEEE Geosci. Remote Sensing Lett., Vol: 12, pp. 1387 – 1391.

5. Renjie Chen; Ning Sun; Xiaojun Chen; Min Yang; Qingyao Wu, Year: 2018, "Supervised Feature Selection With a Stratified Feature Weighting Method", IEEE Access, Vol: 6, pp. 15087 – 15098.

6. Alberto Palacios Pawlovsky, Year: 2018, "An ensemble based on distances for a kNN method for heart disease diagnosis", in 2018 International Conference on Electronics, Information, and Communication (ICEIC), pp. 1 – 4.

7. Pahulpreet Singh Kohli; Shriya Arora, Year: 2018, "Application of Machine Learning in Disease Prediction", in 2018 4th International Conference on Computing Communication and Automation (ICCCA), pp. 1 – 4.

8. Syed Muhammad Saqlain; Muhammad Sher; Faiz Ali Shah; Imran Khan; Muhammad Usman Ashraf; Muhammad Awais; Anwar Ghani, Year: 2018, "Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines", Knowl Inf Syst, Vol: 58, no: 1, pp. 139 – 167.

9. R. Kannan; V. Vasanthi, Year: 2019, "Machine Learning Algorithms with ROC Curve for Predicting and Diagnosing the Heart Disease", in Soft Computing and Medical Bioinformatics, Singapore: Springer Singapore, 2019, pp. 63 – 72.

10. C. Beulah Christalin Latha; S. Carolin Jeeva, Year: 2019, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", Informatics in Medicine Unlocked, Vol: 16, pp. 100203.

11. Saba Bashir; Zain Sikander Khan; Farhan Hassan Khan; Aitzaz Anjum; Khurram Bashir, Year: 2019, "Improving Heart Disease Prediction Using Feature Selection Approaches", in 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, pp. 619 – 623.

12. Nabaouia Louridi; Meryem Amar; Bouabid El Ouahidi, Year: 2019, "Identification of Cardiovascular Diseases Using Machine Learning", in 2019 7th Mediterranean Congress of Telecommunications (CMT), Fès, Morocco, pp. 1 – 6.

13. K. Sathya; R. Karthiban, Year: 2020, "Performance Analysis of Heart Disease Classification for Computer Diagnosis System", in 2020 International Conference on Computer Communication and Informatics (ICCCI), pp. 1 – 7.

14. A. Singh; R. Kumar, Year: 2020, "Heart Disease Prediction Using Machine Learning Algorithms", in 2020 International Conference on Electrical and Electronics Engineering (ICE3), pp. 452 – 457.