A Comprehensive Survey on Heart Disease Prediction

Varsha Singh [1], Ankit Arora [2]

[1] M.Tech. Scholar, [2] Assistant Professor

Department of Computer Science and Engineering, Vishwavidyalaya Engineering College,
Ambikapur, Chhattisgarh, India

Abstract

Heart disease, stroke, and other vascular illnesses will kill one-third of humanity, according to the World Health Organization. Reducing mortality rates and providing the best clinical decision support for cardiac patients necessitates the use of an appropriate machine learning model to target early detection and accurate heart disease prediction. In this study, the Cleveland and Z-Alizadeh Sani datasets are used to examine the efficacy of several heart disease prediction algorithms.

Keywords: Coronary Heart Disease, Heart Disease Prediction, Clinical Decision Support

1. Introduction

The complex and sometimes deadly nature of heart disease (HD) has long been known. This condition causes abnormal cardiac function, which in turn blocks blood arteries and increases the risk of heart attack, angina, and stroke. Coronary artery disease, coronary heart disease, congestive heart failure, and abnormal cardiac rhythms are the most frequent forms of heart disease. Conventional risk variables such as age, sex, hypertension, excessive cholesterol, irregular pulse, and many others provide significant difficulties in early prediction of such HD [1]. Even in economically deprived regions and rural areas, cardiovascular disease (CVD) has been recognised as one of the leading causes of mortality in India. This is despite the fact that cardiovascular risk factors vary widely across various segments of society. The primary impetus for this research came from worldwide data showing that premature death due to CVD has increased from 23.2 million in 1990 to 37 million in 2010 at an annualised rate of 59% [2].

Because of the importance of getting an accurate diagnosis of heart disease, a number of invasive clinical procedures have been developed, such as the angiography. This has prompted a number of scientists to investigate the feasibility of using data-mining methods for the reliable diagnosis of CVD.

What we mean by "Machine Intelligence" is the ability of machines to learn and adapt, allowing them to solve problems and collaborate with other machines and the physical environment [2].

Artificial Intelligence (AI) techniques like machine learning and deep learning will likely serve as the basis for the model used to make predictions and verify the data. Both are very effective and deserve to be used in medical data analytics. Consider using several machine intelligence paradigms if you are looking for a reliable method for diagnosing a cardiac disease that will also help with prediction, monitoring, and other clinical management tasks.

In the following sections, we will delve deeply into the relevant works of machine / deep learning in the medical area of heart disease predictions, and we will also look at the generic framework picked by most researchers for the prediction of heart illnesses (Figure 1). In what follows, we provide a primer on the heart disease datasets that researchers often use.

In the findings portion of this article, we provide the positive and negative aspects of the evaluated articles, and in the discussion section, we focus on the most important issues.

2. The Heart Disease Data Sets

In this part, we present a brief summary of the most prevalent types of datasets used by the articles under consideration.

Researchers rely heavily on the Cleveland heart disease dataset for machine learning, which may be accessed via UCI's digital repository. There are a total of 303 samples, with 6 having unknown values. Although there are 76 characteristics in the raw data, only around a dozen will be included in the published paper, and one of them will describe the impact of the condition.

Researchers also often use the Z-Alizadeh Sani dataset, which consists of data from 303 patients and has 55 input factors and a class label variable for each. There are four subtypes of coronary heart disease represented by the class label variable: normal, LAD, LCX, and RCA. The primary goal of collecting this data was to aid in the detection of CAD. Table 2 explains the attributes and their acceptable ranges.

StatLog Heart, the Hungarian, the Long Beach VA, and the Kaggle Framingham dataset are also employed in the prediction method by the researchers. Table 1 shows that there are 13 traits shared by the 270 samples that make up the Statlog dataset and that these features are comparable to those found in Cleveland.

In contrast to the Cleveland dataset shown in Table 1, the Hungarian and Long Beach VA datasets are available from the UCI repository, and both include 274 samples for each of the 14 characteristics.

The Kaggle Framingham dataset has a vast amount of information, with samples totalling 4,240 patients and 16 characteristics that integrate behavioural, demographic, and medical risk variables.

When this process fails, a lump of tissue called a tumor results, that is, when the former cells are left behind and the young cells expand needlessly. New cells are created and old ones are destroyed in a healthy human body.

The WHO has noted that the growing radio frequency electromagnetic field connected to electronics devices such as cell phones may be the cause of brain tumors. Tumors are a deadly illness, according to the National Health Portal, Government of India, with a survival rate of less than 4% for surviving for greater than 4 years.

Different methods, such as neurological testing, angiograms, spinal taps, CT scans, and MRIs, aid in the identification of brain tumors based on symptoms and family history.

We looked at a number of newly popular strategies in this research to segment and categorize the brain tumor seen in MRI images. We have also provided a comparison based on how well the techniques for classifying abnormality and normalcy have worked. We also spoke about the brain tumor datasets that are currently available for future technique validation.

3. Basic Prediction

The following figures shows the basic diagram of a heart disease prediction system.

Figure 1: Basic Operational Components

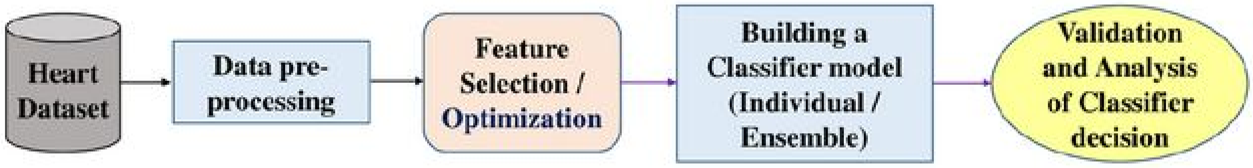
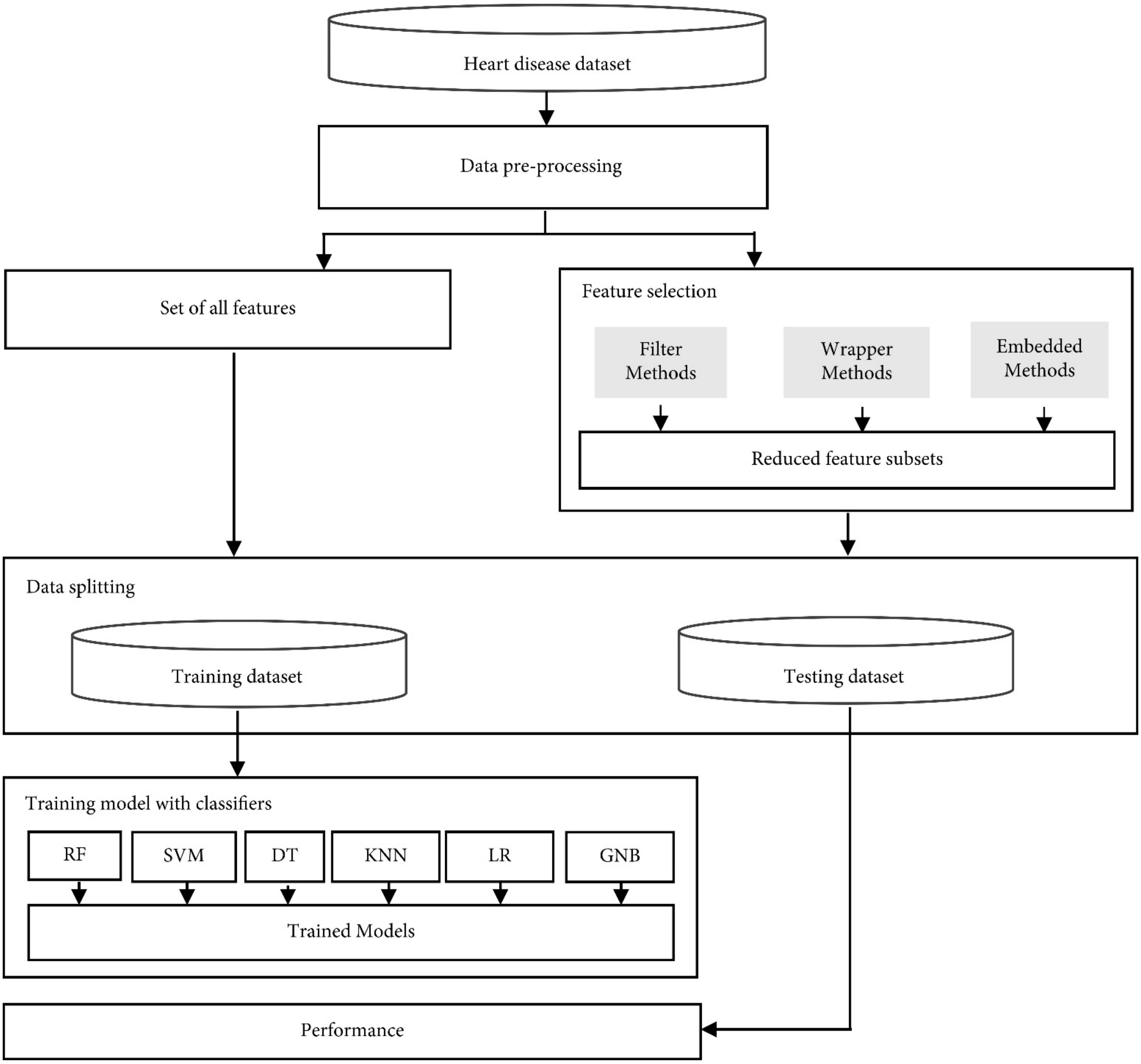

Heart Dataset
Data pre-processing
Feature Selection / Optimization
Building a Classifier model (Individual / Ensemble)
Validation and Analysis of Classifier decision

Figure 2: Framework of Prediction



4. Comparison and Results

The comparative assessment of numerous evaluated publications related to heart disease prediction are shown in Table 1.

Table 1: Comparative Study of the Popular Works

Reference No.	Predicted Cardiac Disease	Features	Data Size
3	Cardiac arrest	System to predict cardiac arrest within 72 hours	1386 real patient data
4	Heart disease	Using artificial intelligence to diagnose	Cleveland dataset

	(General)	heart disease	
6	Arrhythmia diagnosis	Identify 5 heartbeat types automatically.	Cleveland dataset
7	Myocardial infarction (MI)	Data mining for MI prediction	Real data of 455 healthy and 295 MI cases
9	Heart failure	30-day readmission or death from heart failure prediction	Z-Alizadeh Sani dataset
12	Coronary Heart Disease	Determine substantial class imbalances in clinical data classification.	Z-Alizadeh Sani dataset
15	Cardiovascular diseases	Recognizing the characteristics of the handcraft in a precise manner	MIT-BIH dataset
16	Myocardial Infarction	To identify and localize myocardial infarction	Cleveland dataset
19	Chronic heart failure (CHF)	Explored the incremental benefits that might be obtained from each characteristic	Based on 487 real patient data
20	Arrhythmia	Multi-class Arrhythmia detection employing hybrid spatial-temporal features	MIT-BIH

5. Discussion

Over the last few years, a number of researchers have put a significant amount of effort into developing methods to forecast instances of heart disease using the datasets described above.

In 1979, GA Diamond and J.S. Forrester used Bayes' Theorem [5] to draw a diagnostic conclusion regarding the likelihood of illness in a particular patient based on data from procedures such as stress electrocardiography, cardio kymography, thallium scintigraphy, and cardiac fluoroscopy. This conclusion was based on data from these procedures and was used to draw a conclusion regarding the likelihood of illness in the patient. Later on, W.F. Wilson et al. [6] introduced a new facet to the procedures for heart disease by predicting CHD based on risk factor categories using regression equations and logistic methods. This was done in an effort to improve the accuracy of the methodologies.

During the later stages of the project, a number of academics collaborate to develop novel machine learning and deep learning algorithms in order to provide accurate projections about the occurrence of cardiovascular sickness.

This article is a synopsis of current studies conducted on determining the likelihood of developing cardiovascular disease.

A profusion of research on the prognosis of cardiac disease have been carried out in the recent years by a variety of academics making use of the datasets described above.

As early as 1979, G.A. Diamond and J.S. Forrester consolidated data from multiple tests, including stress combining the findings of ECG, CK, thallium scintigraphy, and cardiac fluoroscopy into a single diagnosis. This was done by stress combining the findings of ECG, CK, thallium scintigraphy, and cardiac fluoroscopy.

The chance of illness in a particular patient may be determined using Bayes' Theorem [5]. In following years, cardiology made some progress in its attempt to quantify CHD by classifying probable risk factors. This was an important step forward using logistic and regression techniques developed by W.F. Wilson and colleagues [6]. Many researchers then develop machine learning and deep learning methods in the subsequent phases using the datasets from the UCI repository in order to predict cardiovascular disease [7-18].

The prognosis of cardiovascular sickness is the topic of discussion in this research, which includes a literature review of works on the subject.

There were many instances in which the accuracies were found to be higher than expected, based on the features and machine learning algorithms that were used. When it came to reaching a high level of accuracy, a number of models recommended employing a limited sample size rather than a factor that was highly associated, such as age [7]. In contrast to [10], the research that was published in [9], [11], and [12] provides high accuracy with complete features due to the use of a technique that is both efficient and compact. A step nearer, on the other hand, a more in-depth examination of [10] reveals that although using the identical ML approach, it has a lower degree of accuracy than [12]. To put it another way, this illustrates that the size of the sample is quite important for determining the trustworthiness of predictions.

Some researchers choose for other approaches, such as feature selection and optimization procedures, in order to improve the accuracy of their predictions. These tactics remove data that have lower correlations. For instance, the removal of one or more strongly correlated and necessary parameters for the diagnosis of the illness, such as age, resting ECG, ST Depression, etc., leads to improved accuracy, as demonstrated in a number of representations [18, 20, 21]. These parameters include things like age and resting ECG.

However, the practise of feature selection in prediction models [14, 16, 17, 19] has not only increased accuracy but also mitigated problems such increased processing costs and overfitting brought on by irrelevant input features during the learning process. In addition, the approaches may also present problems with the design, which may be solved with the assistance of the most appropriate advanced prediction models within the framework of a prospective research project.

There are several techniques to assess the effectiveness of segmentation or classification systems. Researchers demonstrate their verified findings using a variety of approaches. Mean Square Error (MSE), Confusion matrix, Jaccard Index, Peak Signal to Noise Ratio (PSNR), Specificity, Accuracy metric, Recall, Sensitivity, and Precision are some of the commonly used performance measures that are analyzed in this study. The crucial information regarding the actual outcome and the projected outcome given by segmentation or classification algorithms is provided by confusion matrices.

6. Conclusion

Machine intelligence may be used as an alternative diagnostic approach to forecast sickness and keep patients informed.

This article examines machine learning, ensemble, and deep learning cardiac prediction systems. From the studied literature, the Cleveland heart disease dataset with 303 cases and 14 characteristics is most utilized. Small sample sizes are to blame. Any research using additional data sources used a single dataset with few characteristics. As a result, high-accuracy prediction models produced by removing extraneous information, removing strongly correlated components, or employing feature selection / optimization approaches cannot be generalized, which is a severe flaw.

Despite the researcher's efforts, prediction models are not standardized. Investigate alternative heart disease datasets with more characteristics to improve classification and prediction accuracy. Future studies will focus on developing a predictive framework model that addresses most of this paper's flaws. In addition, real-time data should be analysed using the working learning model to verify clinical correlation and validation.

References

1. Cardiovascular Diseases (CVDs), "World health organization". https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [last accessed on 12 August 2022]
2. K. Srivastava, D.K. Choubey, "Heart disease prediction using machine learning and data mining", International Journal of Recent Technology and Engineering, vol. 9, no. 1, pp. 212–219, 2020.
3. R. Aggrawal, S. Pal, "Sequential feature selection and machine learning algorithm-based patient's death events prediction and diagnosis in heart disease", SN Computer Science, vol. 1, no. 6, 2020.
4. X.-Y. Gao, A.A. Ali, H.S. Hassan, E.M. Anwar, "Improving the accuracy for analyzing heart diseases prediction based on the ensemble method", Complexity, vol. 2021, Article ID 6663455, 2021.
5. E.M. Senan, M.H. Al-Adhaileh, F.W. Alsaade et al., "Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques", Journal of Healthcare Engineering, vol. 2021, Article ID 1004767, 2021.
6. J.P. Li, A.U. Haq, S.U. Din, J. Khan, A. Khan, A. Saboor, "Heart disease identification method using machine learning classification in e-healthcare", IEEE, vol. 8, 2020.
7. G. Angayarkanni, "Selection of features associated with coronary artery diseases (CAD) using feature selection techniques", Journal of Xi'an University of Architecture & Technology, pp. 686–689, 2020.
8. A. Dutta, T. Batabyal, M. Basu, S.T. Acton, "An efficient convolutional neural network for coronary heart disease prediction", Expert Systems with Applications, vol. 159, Article ID 113408, 2020.

9. A.K. Dwivedi, S.A. Imtiaz, E.R. Villegas, “Algorithms for automatic analysis and classification of heart sounds - A systematic review”, IEEE Access, vol. 7, 2019.
10. L.A. Allen, L.W. Stevenson, K.L. Grady et al., “Decision making in advanced heart failure: A scientific statement from the American heart association”, Circulation, vol. 125, no. 15, pp. 1928–1952, 2012.
11. S. Ghwanmeh, A. Mohammad, A. Al-Ibrahim, “Innovative artificial neural networks-based decision support system for heart diseases diagnosis”, Journal of Intelligent Learning Systems and Applications, vol. 5, no. 3, pp. 176–183, 2013.
12. Jesmin Nahar, Tasadduq Imam, Kevin S. Tickle, Yi-Ping Phoebe Chen, “Computational intelligence for heart disease diagnosis: A medical knowledge driven approach”, Expert Systems with Applications, 40(1), 96–104, 2013.
13. Houda Benhar, Ali Idri, J.L. Fernandez-Aleman, “Data preprocessing for heart disease classification: A systematic literature review”, Computer Methods and Programs in Biomedicine, 2020.
14. Adyasha Rath, Debahuti Mishra, Ganapati Panda, Suresh Chandra Satapathy, “An exhaustive review of machine and deep learning based diagnosis of heart diseases”, Multimedia Tools and Applications, 1–59, 2021.
15. Simran Verma, Abhishek Gupta, “Effective prediction of heart disease using data mining and machine learning: A review”, 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 249–253, 2021.
16. Narender Kumar, Dharmender Kumar, “Machine learning based heart disease diagnosis using non-invasive methods: A review”, Journal of Physics: Conference Series, volume 1950. IOP Publishing, 2021.
17. Taissir Fekih Romdhane, Mohamed Atri Pr. “Electrocardiogram heartbeat classification based on a deep convolutional neural network and focal loss”, Computers in Biology and Medicine, 123, 2020.
18. Mikkili Dileep Kumar, K.V. Ramana, “Cardiovascular disease prognosis and severity analysis using hybrid heuristic methods”, Multimedia Tools and Applications, 80(5), 7939–7965, 2021.
19. Yilin Wang, Le Sun, Sudha Subramani, “Cab: Classifying arrhythmias based on imbalanced sensor data”, KSII Transactions on Internet and Information Systems (TIIS), 15(7), 2304–2320, 2021.
20. Issam Salman, “Heart attack mortality prediction: An application of machine learning methods”, Turkish Journal of Electrical Engineering & Computer Sciences, 27(6), 4378–4389, 2019.
21. Dafni K. Plati, Evanthia E. Tripoliti, Aris Bechlioulis, Aidonis Rammos, Iliada Dimou, Lampros Lakkas, Chris Watson, Ken McDonald, Mark Ledwidge, Rebabonye Pharithi et al., “A machine learning approach for chronic heart failure diagnosis”, Diagnostics, 11(10), 1863, 2021.