Machine Learning in Cardiology: A Survey of Early Detection Models for Heart Diseases

Mudasir Ahad[1], Devanand Padha[2], Himanshu Sharma[3]

[1]M. Tech Student, Department of Computer Science and Information Technology,
Central University of Jammu, Samba, India, 181143
[2]Professor, Department of Computer Science and Information Technology,
Central University of Jammu, Samba, India, 181143
[3]Research Scholar, Department of Computer Science and Information Technology,
Central University of Jammu, Samba, India, 181143

Abstract: Heart disease detection and early prediction is one of the most difficult tasks in the medical field. Almost two people die every minute due to cardio vesicular diseases. According to World Health Organization (WHO), 17.9 million people depart their life every year out of which 4.77 million people are from India alone. About 13% of the world's total population is involved in cardiac disease. Early detection of the disease is crucial for effective treatment that can save millions of lives in the world. Traditional methods of heart disease detection typically involve a combination of medical history, physical examination, and diagnostic tests which are less accurate. With the advancements in machine learning and deep learning techniques, the development of accurate prediction models for heart disease has become possible. Nowadays a large volume of data is being generated in the healthcare sector, which can be leveraged to empower the development of accurate prediction models for heart diseases. Various techniques such as logistic regression, decision trees, random forest, support vector machine, artificial neural networks, and convolutional neural networks have been applied to predict heart diseases. Over the years, advancements in medical technology have led to the development of new diagnostic tools and techniques for detecting heart disease. In this study, a comparative analysis of these techniques is carried out to understand the architectures, parametric characteristics, and datasets involved in heart disease prediction. Our analysis indicates that most heart disease prediction system that have been designed using deep learning algorithms show promising performance.

Keywords: Decision Tree, Naive Bayes, Logistic Regression, Random Forest, Heart Disease Prediction

1. Introduction

Heart diseases are a leading cause of death worldwide and their early detection and management are critical for improving a patient’s health and other body components, such as the brain, kidney, etc. Heart disease is a condition that impairs the heart's ability to pump blood. The signs and symptoms of heart disease can vary depending on the individual. They frequently experience breathlessness, chest pain, arm and shoulder pain, back pain, jaw pain, neck pain, and stomach issues. cardiac failure, and stroke, are a few of the frequent cardiac conditions that affect the heart and blood vessels, including coronary artery disease, heart failure, arrhythmias, and congenital heart defects. Cardiomyopathy is one of several dangerous diseases that receive a great deal of interest in medical research. Because predicting heart attacks can be a complex undertaking requiring much knowledge and expertise, it might be challenging for doctors to do so. The risk factors for heart disease include lifestyle factors such as smoking, lack of physical activity, poor diet, and excessive alcohol consumption. Other risk factors include high blood pressure, high cholesterol, diabetes, obesity, and a family history of heart disease. Diagnosis of heart disease typically involves a combination of medical history, physical examination, and diagnostic tests such as electrocardiogram (ECG), echocardiogram, cardiac catheterization, and stress tests.

Early detection of heart disease is significant because it allows for timely intervention, which can help prevent or minimize the damage caused by the disease. If heart disease is detected early, lifestyle changes such as a healthy diet, regular exercise, and quitting smoking can help manage the disease and prevent further progression which can save millions of lives throughout the world.

Machine-driven predictions about a patient's heart state can be made using a heart diagnostic. It's important to pay attention to the patient's physical exam, symptoms, and indicators. These conditions can lead to a range of symptoms, including chest pain or discomfort, shortness of breath, fatigue, dizziness, and palpitations. While traditional risk assessment methods rely on individual risk factors and clinical data, machine learning (ML) models offer the potential to leverage large and diverse datasets to develop more accurate and personalized predictions of heart disease risk. The use of ML in heart disease prediction is a rapidly evolving field, with numerous studies demonstrating the effectiveness of various ML algorithms in predicting heart disease risk and improving clinical decision-making. Novel approaches to heart disease prediction using ML include the integration of diverse data sources, such as wearable devices and environmental data, to improve risk assessment and the use of explainable Artificial intelligence (AI) techniques to identify the most relevant risk factors and potential interventions. Furthermore, the development of continuous learning models that can adapt to changing patient conditions and incorporate new data over time has the potential to further improve the accuracy and effectiveness of heart disease prediction. As ML models continue to evolve, there is great potential for their use in population health management, personalized medicine, and point-of-care decision support to improve patient outcomes and reduce the burden of heart disease.

With the rapid expansion of the internet, medical progress, and the emergence of pandemics like Covid-19, more data has been generated in the medical sector. This data is generated from various sources such as electronic health records, medical imaging, genetic testing, and social media platforms. During pandemics patients were unable to visit the doctor, especially old age patients, we need to develop a remote assistance system for pandemic situations which is important to help ensure the safety and well-being of individuals, especially those who are at high risk for the disease. Researchers are working on developing a system that will be easy for the patients for diagnosis of heart diseases. Despite the widespread interest in the field, only a few survey studies have been published. Although prior studies have covered a thorough comparative assessment of the heart disease prediction framework, these surveys could only cover a subset of heart disease prediction frameworks as most of the models have been developed recently. As a result, we propose to fill the above-discussed research gaps by undertaking this comprehensive review of heart diseases converting both machine learning and deep learning-based techniques along with their datasets in this single duty. The following are the contributions of our survey study:

i. We review the existing literature on heart disease prediction frameworks using our novel proposed taxonomy.

ii. We perform a comparative evaluation of the datasets used in heart disease prediction.

iii. We identify a list of open research challenges and future scopes in the field of heart disease perfections.

The rest of the article is organized as follows. Section 2 discusses the basic architecture and our proposed taxonomy of the heart disease precision framework. Section 3 describes a comprehensive literature on the heart disease prediction framework. Section 4 evaluates the existing datasets used in heart disease prediction. Sections 5 and 6 sketch out the open research challenges, future directions, and conclusions. The abbreviations used in this article are tabulated in Table 1.

Table 1: A list of abbreviations used in this article with their meanings

Abbreviation	Meaning
ML	Machine Learning
AI	Artificial intelligence
ANN	Artificial Neural Network
SVM	Support Vector Machine
DT	Decision Tree
DNN	Deep Neural Network
CNN	Convolutional Neural Network
WHO	World Health Organization
CAD	Coronary Artery Disease

2. Heart disease prediction framework

A heart disease prediction framework is a system that uses various data analysis and machine learning techniques to predict the likelihood of a person developing heart disease. The framework typically consists of two components feature extractor and discriminator as shown in figure 1. The input to this model is a series of parametric attributes of patients concerning their age, sex, cholesterol, etc. This data is fed to a deep neural network that tries to decode the deep feature correlations between the data. Based

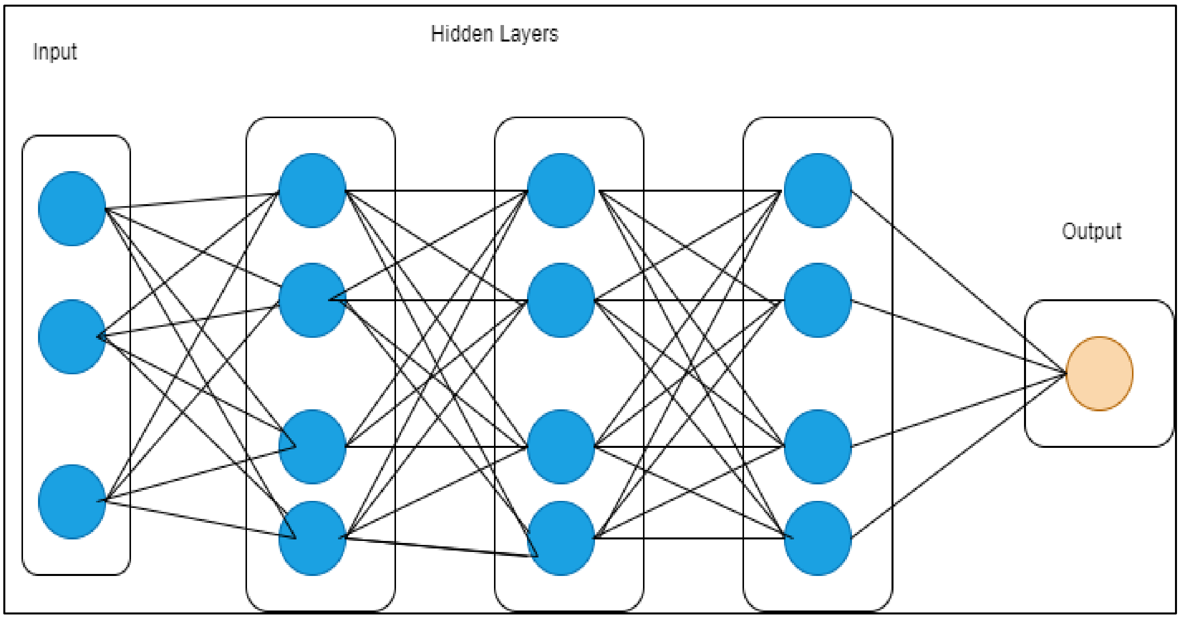

Figure 1: A generic architecture of heart disease prediction framework

on this correlation analysis, the final prediction about whether the patient is infected or not is made. A typical heart disease prediction framework thus consists of a pipeline of feature extractor and discriminator. Both subcomponents are discussed below.

I. Feature Extractor: A feature extractor is a component of a machine learning system that is responsible for identifying and extracting relevant features from input data that can be used to make predictions or classifications. In the context of heart disease prediction, feature extraction involves identifying specific factors or variables that are strongly associated with the likelihood of developing heart disease.

II. Discriminator: A discriminator is a component of a machine learning system that is responsible for distinguishing between different classes or categories of data. In the context of heart disease prediction, a discriminator is a machine learning algorithm that is trained to differentiate between individuals who are likely to develop heart disease and those who are not.

Based on the type of feature extraction and discrimination model being used, the heart disease prediction frameworks can be classified into two abstract types namely machine learning and deep learning-based heart disease prediction models shown in Figure 3.

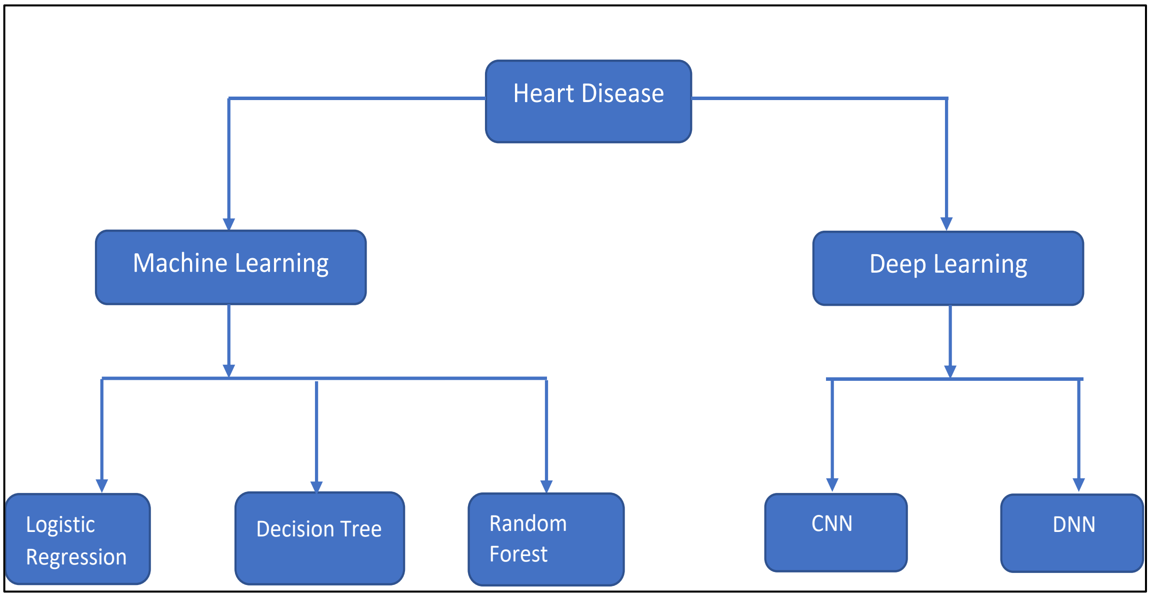

Figure 3: Taxonomy of heart disease prediction frameworks

Machine learning is a subfield of artificial intelligence that involves the development of algorithms and models that enable computers to learn from data and improve their performance on a specific task without being explicitly programmed. Machine learning has been used extensively in the prediction and diagnosis of heart disease. Heart disease is a complex and multifactorial condition, and machine learning algorithms are particularly well-suited to handle the large and complex datasets involved in heart disease prediction. The following machine learning models are employed in the heat disease prediction pipeline.

I. Logistic regression: Logistic Regression is a statistical method and a type of generalized linear model (GLM) that is used for predicting binary outcomes (i.e., outcomes with two possible values, such as success/failure or true/false). The logistic regression model is trained using labeled data, and the goal is to learn a set of weights that can be used to make predictions on new, unseen data. The predictions are made by applying the trained model to the new data, which results in a probability value between 0 and 1 that represents the likelihood that the outcome is positive (i.e., success or true).

II. Decision Tree: A decision tree is a machine learning algorithm that is used for classification and regression tasks. It is a simple but powerful algorithm that is widely used for predictive modeling in various fields, including healthcare. The decision tree algorithm creates a tree-like model of decisions and their possible consequences, represented by branches and nodes. The tree starts with a single node, which represents the entire dataset, and then recursively splits the dataset into smaller subsets based on the most important features. The goal of the algorithm is to find the optimal splits that maximize the separation between the classes or minimize the error in the regression task.

III. Random Forest**:** Random Forest is an ensemble learning method for classification and regression that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The idea behind random forests is to combine the predictions of multiple decision trees, which reduces the risk of overfitting and improves the overall performance of the model. The random forest algorithm creates a set of decision trees from randomly selected subsets of the training data, which are used to make predictions.

IV. Naive Bayes: Naive Bayes is a probabilistic machine learning algorithm that is commonly used for classification tasks. It is a simple yet effective algorithm that is widely used in various fields, including healthcare. The Naive Bayes algorithm is based on Bayes' theorem, which is a statistical principle that describes the probability of an event occurring given some evidence or information. In the case of classification, the algorithm uses Bayes' theorem to calculate the probability of an input belonging to a particular class, given the observed features or predictors.

Deep learning is a type of machine learning that involves the use of neural networks with multiple layers. Neural networks are composed of interconnected nodes or neurons that perform simple computations on the input data and then pass the results to the next layer of neurons for further processing. The following machine learning models are employed in the heat disease prediction pipeline.

I. Convolutional Neural Network (CNN): A Convolutional Neural Network (CNN) is a type of deep neural network that is commonly used for image recognition and computer vision tasks. It is particularly effective in analyzing images because it can automatically learn hierarchical features or patterns from the data, which are essential for image understanding. CNNs are composed of multiple layers, including convolutional layers, pooling layers, and fully connected layers. The convolutional layers are responsible for learning feature maps by applying a set of filters or kernels to the input image. The pooling layers downsample the feature maps to reduce their spatial dimensionality, while the fully connected layers perform classification or regression on the learned features.

II. Deep Neural Network (DNN): Deep Neural Networks (DNNs) are a type of artificial neural network (ANN) that is composed of multiple layers of interconnected artificial neurons. These networks are designed to mimic the behavior of biological neurons and can perform complex computations. DNNs are commonly used for tasks such as image and speech recognition, natural language processing, and computer vision. They have been shown to achieve state-of-the-art results on a wide variety of tasks and are a key component of many artificial intelligence systems.

3. Literature Survey

Year	Author	Deep Learning	Technique	Accuracy
2020	Pandita, A [1]	NO	Logistic regression	84.58%
		NO	KNN	89.71%
		NO	Naïve bayes	79.12%
		NO	Decision Tree	78.90%
2020	Rajdhan, [2]	NO	Logistic regression	82.12%
		NO	Decision Tree	80.79%
		NO	Random Forest	90.00%
		NO	Naïve bayes	84.59%
2020	Ravindhar et [3]	NO	Naïve bayes	98.91%
		NO	K-Nearest Neighbour	87.51%
		NO	Random Forest	89.96%
2021	Ramalingam et al [4]	NO	K-Nearest Neighbour	80.25%
		NO	Logistic regression	87.6%
		NO	Random Forest	84.52%
2021	Jindal et al. 2021 [5]	NO	Logistic regression	83.87%
		NO	K-Nearest Neighbour	87.88%
		NO	Support Vector Machine	86.86%
		NO	Naïve Bayes	84.85%
		NO	Random Forest	86.86%
2021	Rahman et al. 2019 [6]	No	Standard Linear Model	86.76
		NO	Decision Tree	79.78%
		NO	Random Forest	87.64%
		NO	Support Vector Machine	86.52%
		YES	Neural Network	93.03

To predict cardiac diseases, Ravendra et, al [3] used the machine learning algorithms KNN and Random Forest. Post-data collection and data analysis, a link between the various variables and how they affected the desired value was found. The resulting UCI dataset was made accessible on Kaggle. It was divided into two halves for training and testing, 80:20.

Table 2: A comparison of several heart disease prediction articles utilizing machine learning models

Chest discomfort and the highest heart rate reached was shown to positively connect with the goal characteristic. This model was 81.967% accurate while using Random Forest and 86.885% accurate when using KNN. Ramalingam et al.[4] Proposed a web-based application prediction model that splits the UCI dataset in half for training and testing (75–25). The most accurate prediction models were found to be those based on Logistic Regression, with an accuracy rate of 82.89%, followed by SVM (81.57%), Naive Bayes (80.43%), and Decision Tree (81.57%). The user can utilize the online tool as a preliminary examination to evaluate their cardiac health and, if necessary, seek medical advice. According to Rajdhan et al.[2] The system uses four classification algorithms to predict the patient's health, including Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), and Naive Bayes(NB) . Training data and testing data are the two categories of data, respectively. The decision was made to create a confusion matrix that would display both true and false positives in addition to true and false negatives. The Random Forest classification method's maximum accuracy was 90.16%. A set of models employing supervised learning approaches using the WEKA tool was challenged by Devansh Shah et al. [14]. Four distinct classification methods NB, KNN, RF, and DT were used to predict the chance of developing a heart ailment. Before being integrated and reduced, the dataset was cleaned, smoothed, normalized, and aggregated. The KNN technique yielded the most accurate results. proposed in Jindal et al. [5] three separate classification algorithms KNN, RF, and LR—were combined to produce a system with an accuracy rate of 87.5% . Logistic Regression and KNN outperform RF, or an efficient method for predicting heart disease, with K-Nearest Neighbor having the highest accuracy of the three algorithms (88.52%). To create a web application that accepts patients' medical information and calculates whether or not they have a cardiac ailment, Pandita, et al [1] suggested a prediction standard that contains 5 machine learning techniques that utilize the approach with absolute precision. A Flask-based framework and HTML/CSS were used to create the web application. KNN had the highest accuracy, at 89.06%, while Logistic Regression had the lowest accuracy, at 84.38%.

A strategy for anticipating heart illness that is time and money efficient and employs a web application was developed by Saranya et al [7]. The model employs the Random Forest and KNN methods. The dataset from one of Coimbatore's hospitals produced accuracy of 100% and 91.36% using Random Forest and KNN after cleaning and preprocessing. Additionally, with accuracy rates of 98.77% and 95.06%, the probabilities are predicted using an ensemble model with and without logistic regression. The model developed by Akella & Akella [9] The Neural Networks model had the greatest accuracy (93.03%)and recall (93.8%) out of the six predictive models that were applied to the UCI dataset, both of which suggest minimal odds of false negatives and, as a result, incredibly exact findings. The accuracy of the other five models ranged from almost 80% to higher. Ravendra et al [8] proposed backpropagation neural networks, fuzzy KNN, naive Bayes, logistic regression, and k-means clustering were the five methods used. In the experimental investigation of cardiac problems, the 10-fold cross-validation approach is employed. Data was acquired with a maximum accuracy of 98.2%, 87.64% recall, and 89.65% precision using back propagation neural networks. A comparative analysis of all the heart disease predictions is presented in Table 2.

4. Open Research Challenges

There are several research challenges in the field of heart disease prediction. Some of the main areas of focus include:

I. Limited accuracy in current risk prediction models: Current models for predicting heart disease risk are based on a limited set of risk factors and may not accurately predict risk for all individuals. There is a need for more comprehensive models that consider a wider range of risk factors, such as genetic and lifestyle factors.

II. Lack of ensembled modeling approaches: Machine learning and deep learning techniques have the potential to improve the accuracy of heart disease prediction by analyzing large, complex data sets. However, more research is needed to develop and validate these methods.

III. Less use of multimodal data: heart disease is a complex disease and there is limited research on using multimodal data such as clinical data, imaging data, and physiological data in machine learning models to improve the prediction performance.

IV. The severity of robust Dataset: Using a limited dataset in heart disease prediction systems can result in a less accurate model, which may lead to misdiagnosis or incorrect prediction of the disease. To develop a robust and accurate heart disease prediction model, it is important to have a dataset that includes enough samples and a wide range of features that are relevant to heart disease.

5.Conclusion

The use of machine learning techniques for heart disease prediction has shown great promise in improving patient outcomes and reducing the burden of disease on healthcare systems. Machine learning models can integrate a broad range of data sources and employ sophisticated feature selection and ensemble learning techniques to identify the most relevant predictors of heart disease and improve the accuracy and robustness of the prediction model. However, the development and deployment of machine learning models for heart disease prediction also pose several challenges, including data privacy and security, algorithm bias and interpret ability, and integration with clinical work-flows. Addressing these challenges will require ongoing collaboration between researchers, clinicians, and policymakers to ensure the ethical and responsible use of ML in health care. Despite these challenges, the potential benefits of machine learning in improving heart disease prediction are clear, and continued research and development in this area will be critical to realizing the full potential of personalized and precision medicine in improving cardiovascular health.

References

1. Pandita, A., Yadav, S., Vashisht, S., & Tyagi, A. (2021). Review Paper on Prediction of Heart Disease using Machine Learning Algorithms. International Journal for Research in Applied Science and Engineering Technology, 9(6).
2. Saranya, G., & Pravin, A. (2020). A comprehensive study on disease risk predictions in machine learning. International Journal of Electrical and Computer Engineering, 10(4), 4217.
3. Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. International Journal of Engineering & Technology, 7(2.8), 684-687.
4. Ravindhar N, Anand K, […] Winster International Journal of Innovative Technology and Exploring Engineering (2019) 8(11) 1417-1421.
5. Akella, A., & Akella, S. (2021). Machine learning algorithms for predicting coronary artery disease: efforts toward an open-source solution. Future science OA, 7(6), FSO698.
6. Garg, A., Sharma, B., & Khan, R. (2021). Heart disease prediction using machine learning techniques. In IOP Conference Series: Materials Science and Engineering (Vol. 1022, No. 1, p. 012046). IOP Publishing.
7. Issue 6 www.jetir.org (ISSN-2349-5162) Magar R, Memane R, […] Rupnar P (2020)
8. Rajdhan, A., Agarwal, A., Sai, M., Ravi, D., & Ghuli, P. (2020). Heart disease prediction using machine learning. International Journal of Research and Technology, 9(04), 659-662.
9. Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In IOP conference series: materials science and engineering (Vol. 1022, No. 1, p. 012072). IOP Publishing.
10. Rahman, M. J. U., Sultan, R. I., Mahmud, F., Shawon, A., & Khan, A. (2018, September). Ensemble of multiple models for a robust intelligent heart disease prediction system. In 2018 4th international

conference on electrical engineering and Information & communication technology (ICEEiCT) (pp. 58-63). IEEE.

11. Joo, G., Song, Y., Im, H., & Park, J. (2020). Clinical implication of machine learning in predicting the occurrence of cardiovascular disease using big data (Nationwide Cohort Data in Korea). IEEE Access, 8, 157643-157653.
12. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542-81554.
13. Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. SN Computer Science, 1(6), 1-6.