Advanced Machine Learning Approaches For Predicting Cardiovascular Conditions - A Comprehensive Review

Bisma Sultan, *Aasia Rehman

Aasiya.rehman77@gmail.com

University of Kashmir

Abstract

The primary challenge in the healthcare sector revolves around the timely diagnosis of diseases, a crucial factor in saving numerous lives. Leveraging machine learning classification techniques can profoundly enhance medical practices by delivering precise and rapid disease diagnoses. This not only streamlines the process for both physicians and patients but is especially vital in addressing the global menace of heart disease, the leading cause of mortality. This paper conducts a comprehensive exploration of various machine learning classification techniques designed to assist healthcare professionals in diagnosing heart conditions. We commence with an overview of machine learning, providing concise definitions of commonly employed classification techniques for heart disease diagnosis. Subsequently, we delve into a review of notable research endeavors employing machine learning classification techniques in this domain. Additionally, the paper includes a meticulous tabular comparison of the surveyed studies.

Keywords: *Machine Learning, Heart diseases, heart disease diagnosis; heart disease prediction; machine learning; machine learning classification techniques*

1. Introduction

Artificial intelligence (AI) falls within the realm of computer science and focuses on enhancing the intelligence of computers. Recognizing that learning is a fundamental aspect of intelligence, the subfield of AI known as machine learning (ML) emerged. ML is a swiftly advancing branch of AI with widespread applications, particularly in the healthcare sector. Its significance in healthcare lies in its ability to intelligently analyze the abundance of data within the medical field. The digital revolution has led to the substantial collection and storage of vast amounts of data in recent years. Modern hospitals employ monitoring and data collection devices extensively, generating copious amounts of data daily.

Given the sheer volume of this data, it becomes challenging, if not impossible, for humans to extract meaningful insights. This is where machine learning comes into play, offering a widely adopted solution for analyzing healthcare data and identifying issues. In essence, machine learning algorithms learn from previously diagnosed patient cases. The resulting classifier serves various purposes, including aiding doctors in diagnosing new patients with heightened speed and efficiency. Additionally, it plays a role in training students and non-specialists to diagnose patients [1].

The term cardiovascular disease, commonly known as heart disease, encompasses a range of conditions affecting the heart. According to the World Health Organization, approximately 12 million deaths occur globally each year due to heart disease, making it a leading cause of mortality. Particularly prevalent in developing countries, it claims a significant number of lives, with one person succumbing to it every 34 seconds in the United States alone. India, too, experiences heart disease as a primary cause of death, underscoring its status as a formidable threat to adult lives [2].

Diagnosing heart disease is a pivotal yet challenging task within the healthcare domain, requiring swift, efficient, and accurate assessments to save lives. The diagnostic process involves a battery of tests, with healthcare professionals meticulously scrutinizing the results. Recognizing the critical nature of timely diagnosis, researchers have delved into predicting heart disease, developing various prediction systems employing diverse machine learning algorithms[3]. Some systems have demonstrated superior outcomes compared to others, utilizing renowned datasets like the UCI heart disease dataset for training and testing classifiers, while others leverage data from accessible hospitals.

This survey paper provides an insightful overview of machine learning classification techniques applied in the realm of heart disease diagnosis, shedding light on how previous researchers implemented these techniques. It emphasizes the significance of machine learning in healthcare, showcasing its ability to yield accurate predictions and assist healthcare professionals.

The subsequent sections of the paper are organized as follows: Section 2 covers background topics on machine learning, classification techniques, and the widely used heart disease dataset in this field. Section 3 encompasses a literature review of current research in this area, while Section 4 furnishes a comparative analysis in tabular form, evaluating the classification techniques discussed in Section 3 according to their accuracy. Ultimately, Section 5 encapsulates the conclusion.

2. Background

This section furnishes explanations of pertinent topics addressed in this paper, including machine learning, its techniques with concise descriptions, data preprocessing, performance evaluation metrics, and a brief elucidation of the most commonly utilized heart disease dataset.

2.1. Machine Learning

Machine learning (ML) is a domain of artificial intelligence that involves constructing algorithms that can learn from experience. The way that ML algorithms work is that they detect hidden patterns in the input dataset and build models. Then, they can make accurate predictions for new datasets that are entirely new for the algorithms. This way the machine became more intelligent through learning; so it can identify patterns that are very hard or impossible for humans to detect by themselves. ML algorithms and techniques can operate with large datasets and make decisions and predictions [4]. Figure 1 represents a simplified representation of how machine learning works. In this figure, the dataset, which in our case can be a patient database, is preprocessed first. The preprocessing phase is crucial as it cleans the dataset and prepares it to be used by the machine learning algorithm. The model consists of a single algorithm, or it can contain multiple algorithms working together in a hybrid approach. The output of the model is a classifier; this is where the intelligence is, and this is what will make the prediction. If the classifier receives input data, it can predict without any human interruption. For example, if the dataset that is fed into the model is a medical dataset of healthy and unhealthy patients' information, the input data can be a new patient's information. This input data is entirely new to the classifier and has never been seen before. The classifier will receive this data and will predict whether this new patient is healthy or unhealthy based on past data.

2.2. Methods of Machine Learning

The main ML methods can be classified as follows:

2.2.1. Supervised Learning

In this approach, there is a dataset containing examples along with their corresponding responses (outputs). Through a training process, the algorithm acquires knowledge from the dataset, enabling it to make predictions or responses to new inputs based on its learned information. Classification and regression are instances of the supervised learning technique.

2.2.2. Unsupervised Learning

In this method, the dataset lacks corresponding responses. Consequently, the algorithm endeavors to identify similarities among input values and classifies them according to these similarities. The unsupervised learning technique includes the clustering method.

2.2.3. Reinforcement Learning

This method lies between supervised and unsupervised learning, as the model enhances its performance through interaction with the environment, learning to rectify mistakes. It aims to achieve accurate results through examination and experimentation with various possibilities.

The predominant form of learning is supervised learning, particularly the widely utilized classification technique for prediction. This paper primarily concentrates on studies that employ classification algorithms to diagnose heart disease.

2.3 Categorization of Machine Learning Approaches

Categorization, a subset of supervised machine learning techniques, makes predictions for forthcoming cases using a dataset from the past. In this segment, we offer concise explanations of the most commonly employed classification techniques for predicting heart disease.

2.3.1. Naïve Bayes

The Naive Bayes classifier is a member of the probabilistic classifiers' family grounded in the Naive Bayes theorem. It relies on the strong assumption of independence among features, a fundamental aspect influencing its prediction methodology. Notably straightforward to construct, it generally demonstrates effective performance, rendering it suitable for applications in the medical science domain, particularly in disease diagnosis [5].

2.3.2. Artificial Neural Network (ANN)

This algorithm was developed to imitate the neurons in the human brain. It consists of some nodes or neurons that are connected, and the output of one node is the input of another. Each node receives multiple inputs, but the output is only one value. The Multi-Layer Perceptron (MLP) is a widely used type of ANN, and it consists of an input layer, hidden layers, and an output layer. A different number of neurons are assigned to each layer under different conditions [5].

2.3.3. Radial Basis Function (RBF)

This falls under the category of artificial neural networks (ANN) and shares similarities with the Multi-Layer Perceptron (MLP) Neural Network. However, it differs in terms of the number of hidden layers, approximation technique, number of parameters, and various other factors [5].

2.3.4. Decision Tree

This algorithm exhibits a tree-shaped or flowchart-like configuration, featuring branches, leaves, nodes, and a root node. Attributes are housed in the internal nodes, and the branches indicate the outcomes of tests conducted at each node. Decision Trees (DT) find extensive application in classification tasks due to their ability to function effectively without requiring extensive domain knowledge or parameter tuning [5].

2.3.5. K-Nearest Neighbor (KNN)

This algorithm forecasts the class of a new instance by considering the majority votes from its nearest neighbors. It employs Euclidean distance to determine the proximity of an attribute to its neighbors [5].

2.3.6. Support Vector Machine (SVM)

This algorithm demonstrates valuable accuracy in classification. It is characterized as a finite-dimensional vector space, encompassing a dimension for each feature or attribute of an object [5].

2.4 Data Processing

The effectiveness and precision of the predictive model are influenced not only by the algorithms employed but also by the dataset's quality and preprocessing techniques. Preprocessing involves the procedures applied to the dataset before implementing machine learning algorithms, playing a crucial role in preparing the dataset in a format comprehensible to the algorithm.

Datasets may exhibit errors, missing data, redundancies, noise, and other issues, rendering them unsuitable for direct use by machine learning algorithms. Additionally, the dataset's size can pose challenges, especially when dealing with numerous attributes that make it more challenging for the algorithm to analyze, identify patterns, or make accurate predictions. Addressing such issues involves analyzing the dataset and employing appropriate data preprocessing techniques. These preprocessing steps encompass data cleaning, data transformation, imputation of missing values, data normalization, feature selection, and other measures tailored to the dataset's nature [6].

2.5 Evaluation Metrics

Researchers utilize the following metrics to assess prediction models and demonstrate their performance outcomes. We offer concise definitions for each method without delving into intricate details or mathematical equations.

Accuracy: This metric indicates the percentage of accurate results.

Precision: This metric gauges the relevance of the result.

Recall or Sensitivity: Measures the retrieval of relevant results.

F-Measure: A combination of precision and recall.

Receiver Operation Characteristic (ROC): A graph visualizing the classifier's performance, depicting correctly and incorrectly classified cases [5].

Among these metrics, accuracy is the most widely employed performance evaluation metric in all the research papers discussed in our article. Consequently, this overview article focuses on categorizing, comparing, and reviewing previous work based on accuracy.

2.6 Heart Disease Dataset

The heart disease dataset, widely employed in the majority of research papers, is sourced from the University of California, Irvine (UCI) Center for Machine Learning and Intelligent Systems. It comprises four databases from distinct hospitals, each maintaining 14 features but varying in the number of records. Among these, the Cleveland dataset is most frequently utilized by machine learning researchers, primarily due to its fewer missing attributes and a larger number of records. The "num" field within the dataset denotes the presence of heart disease in patients, taking integer values from 0 (indicating no presence) to 4. The Cleveland dataset encompasses a total of 303 instances.

3. Various Classification Approaches for Heart Disease Prediction

Several researchers employ diverse classification methods to forecast heart disease. In this section, we present a condensed overview of the reviewed literature in this field. We categorized the papers according to the algorithms employed in their predictive models. Many researchers either integrated several algorithms into their studies or conducted comparisons among them. This information is detailed in the final section referred to as the "Hybrid Approach" section.

i. Naive Bayes Approach:

In their study[7], Vembandasamy et al. employed a Naive Bayes classifier to determine the existence or non-existence of heart disease. The dataset utilized in the investigation was sourced from a prominent diabetic research institute in Chennai, encompassing approximately 500 patient records with 11 attributes, including the diagnosis. The Waikato Environment for Knowledge Analysis (WEKA) tool, comprising a set of machine learning algorithms, was employed for the implementation of the Naive Bayes classifier. The accuracy achieved in their study was 86.41%. In their work [8], Medhekar et al. introduced a system that utilized the Naive Bayes classifier to classify data into five categories: no, low, average, high, and very high. The system anticipates the likelihood of heart disease in the input data. The UCI heart disease dataset, as illustrated in Table 1, was employed for both training and testing. The system demonstrated an accuracy rate of 88.96%.

Table 1 Attributes of Dataset

S. No	Attribute	Description
1.	Gender	Male or Female
2.	Age	Age in years
3.	trestbps	Resting blood pressure in mmHg
4.	chol	Serum cholesterol in mg/dl
5.	cp	Chest pain type
6.	thalach	Maximum heart rate achieved
7.	exang	Exercise induced angina
8.	oldpeak	ST depression induced by exercise relative to rest
9.	fbs	Fasting blood sugar
10.	slope	The slope of the peak exercise ST segment
11.	thal	Thallium heart scan
12.	num	Diagnosis of heart disease (angiographic disease status)
13.	restecg	Resting electrocardiographic results
14.	ca	Number of major vessels (0-3) colored by flourosopy

ii. Artificial Neural Network (ANN) Approach:

In their publication [9], Das et al. introduced a system that utilized the Ensemble method with Artificial Neural Network (ANN). The Cleveland heart disease dataset, as depicted in Table 1, served as the basis for their research. The ensemble model enhanced generalization by consolidating multiple models trained on the same task. The experiment was conducted using SAS Enterprise Miner 5.2 as the implementation tool, and the outcomes revealed that the model successfully forecasted heart disease with an accuracy rate of 89.01%. In their study [10], Chen et al. designed a system for predicting heart disease (HDPS) employing Artificial Neural Network. The research utilized Learning Vector Quantization (LVQ), a specific type of ANN. In this article, the Artificial Neural Network (ANN) model employed thirteen neurons for the input layer, six neurons for the hidden layer, and two neurons for the output layer. The dataset utilized in the study corresponds to the Cleveland dataset presented in Table 1. The created system features a user-friendly interface, prompting users to input thirteen medical attributes for accurate predictions. The system output presents the prediction outcome, indicating either a healthy or unhealthy result, along with the ROC curve, accuracy, sensitivity, specificity, and the duration taken to display the results. The system was crafted using the C programming language, with C# utilized for constructing the user interface. The outcomes indicated that the model achieved an accuracy of around 80%, sensitivity of 85%, and specificity of approximately 70%. In their work [11], Dangare and Apte employed Artificial Neural Network (ANN) to create a Heart Disease Prediction System (HDPS) for forecasting the presence or absence of heart disease in patients. The algorithm was trained using the Cleveland heart disease dataset, as indicated in Table 1, and tested with the Statlog dataset. Both datasets were sourced from the UCI repository and consist of thirteen medical attributes. Two more attributes, namely smoking and obesity, were included to enhance accuracy, resulting in a total of fifteen attributes. The WEKA tool was employed for conducting the experiments. The findings indicated that utilizing the thirteen attributes resulted in an accuracy of 99.25%, whereas incorporating the fifteen attributes yielded an accuracy of almost 100% for disease prediction.

iii. Decision Tree Approaches:

In their study [12], Sabarinathan and Sugumaran employed the Decision Tree J48 algorithm for both feature selection and predicting heart disease. The dataset utilized comprises thirteen medical attributes or features, with 240 records employed for training and 120 for testing. The accuracy attained with all features was 75.83%, improving to 76.67% with feature selection. Further removal of irrelevant features increased accuracy to 85%. The study asserts that the J48 algorithm facilitates selecting minimal features to boost prediction accuracy. In their study [13], Patel et al. conducted a comparison of various decision tree algorithms using the WEKA tool on the UCI dataset to ascertain the existence or non-existence of heart disease. Various algorithms, including J48, logistic model tree, and random forest, were assessed. Among them, the J48 algorithm demonstrated superior performance, achieving an accuracy of 56.76%.

iv. K-Nearest Neighbor (KNN) Approaches:

In their work [14], Shouman et al. utilized K-Nearest Neighbor (KNN) for heart disease prediction with the Cleveland dataset. The study compared the outcomes of employing KNN alone and employing KNN in conjunction with the voting technique. Voting involves partitioning the data into subsets and applying the classifier to each subset. The evaluation process is conducted using 10-fold cross-validation. The findings indicated that, in the absence of voting, accuracy varied between 94% and 97.4% across different K values. The highest accuracy, reaching 97.4%, was observed when K was set to 7. However, employing the voting technique did not enhance the accuracy. The outcomes demonstrated that, at K=7, the accuracy decreased to 92.7%.

iv. Support Vector Machine (SVM) Approaches:

In their research[15], Wiharto et al. examined the accuracy of different types of SVM algorithms using the UCI dataset for heart disease diagnosis. The investigation encompassed diverse SVM types, including Binary Tree Support Vector Machine (BTSVM), One-Against-One (OAO), One-Against-All (OAA), Decision Direct Acyclic Graph (DDAG), and Exhaustive Output Error Correction Code (ECOC). The initial step involved preprocessing the dataset with a min-max scaler. Subsequently, the algorithm was trained on the dataset using the SVM algorithms mentioned earlier. During the performance assessment, BTSVM outperformed the other algorithms, achieving an overall accuracy of 61.86%.

v. Hybrid Approaches:

This section comprises studies that constructed models utilizing various algorithms or conducted comparisons among multiple algorithms. Khateeb and Usman in [3] conducted experiments with different classification algorithms, including Naive Bayes, KNN, decision tree, and the bagging technique, using the UCI Cleveland dataset. The study was segmented into six cases, and each classifier calculated accuracy for every case. In Case 1, all classifiers were applied to the dataset without feature reduction. For Case 2, feature reduction was implemented. Rather than utilizing all 14 attributes in the dataset, only seven attributes—deemed the most crucial for heart disease diagnosis—were selected. For Case 3, only the most generic features, such as age, sex, and resting blood sugar, were excluded. In Case 4, the dataset underwent resampling using the WEKA tool, and only the seven most essential attributes were retained. The resampling contributed to an improvement in the accuracy of each classifier. For Case 5, resampling was employed on all 14 attributes. In the concluding Case 6, the Synthetic Minority Over-sampling Technique (SMOTE) was implemented using the WEKA tool. The most favorable outcome was obtained using KNN in Case 5, resulting in an accuracy of 79.20%.

In their study [5], Pouriyeh et al. performed an extensive comparison of various classification techniques using the Cleveland heart disease dataset to identify the classifier that exhibits superior performance. The classifiers considered in the study encompassed Decision Tree (DT), Naive Bayes (NB), Multi-layer Perceptron (MLP), K-Nearest Neighbor (KNN), Single Conjunctive Rule Learner (SCRL), Radial Basis Function (RBF), and Support Vector Machine (SVM). The paper also involved the evaluation of ensemble techniques such as bagging, boosting, and stacking. The authors employed the K-Fold Cross Validation technique to assess the accuracy of classifiers. For each classifier, the assessment metrics included accuracy, precision, recall, F-measure, and ROC curve. The KNN classifier underwent experimentation with various values of K, with K=9 identified as the optimal value. In the case of ANN, various neuron numbers were tested to determine the optimal combination, resulting in thirteen, seven, and two for the input, hidden, and output layers, respectively. The study was split into two experiments: the initial one focused on comparing the various classifiers mentioned earlier, and the second one centered around the application of ensemble techniques. The outcomes indicated that in the first experiment, SVM surpassed the other classifiers with an accuracy of 84.15%. In the second experiment, the employment of the boosting technique with SVM demonstrated the highest efficiency, achieving an accuracy of 84.81%.

In their work [16], Amin et al. introduced a hybrid system for heart disease prediction incorporating both Artificial Neural Network (ANN) and Genetic Algorithm. The dataset employed in the study was obtained from a survey conducted by the American Heart Association, comprising information from 50 individuals and featuring thirteen attributes.Data analysis included preprocessing to eliminate missing or inaccurate values. The dataset was partitioned, allocating 70% for training and 15% each for testing and validation. The system was developed using MATLAB R2012a, utilizing the Global Optimization Toolbox and the Neural Network Toolbox. The outcomes demonstrated an 89% accuracy in predicting the presence or absence of heart disease in individuals.

Waghulde and Patil in [17] created a heart disease prediction system that integrated Artificial Neural Network (ANN) and Genetic Algorithm. The approach involved utilizing a genetic algorithm to initialize the weights in the neural network. The MATLAB experiment utilized a dataset of 50 individuals from the American Health Association, featuring thirteen attributes. With six and ten hidden nodes, the results achieved accuracies of 98% and 84%, respectively. In [18], Amma introduced a heart disease diagnosis system that integrates Artificial Neural Network (ANN) and Genetic Algorithm. The Cleveland dataset was employed, and preprocessing involved filling missing values and normalizing the data using Min-Max normalization. The genetic algorithm was employed to determine the weights of the neural network, resulting in an accuracy of 94.17%.

In their study [19], Venkatalakshmi and Shivsankar incorporated a comparison between Naive Bayes and Decision Tree to identify which one exhibited the highest accuracy for predicting heart disease. The UCI heart disease dataset was utilized, and the experiment conducted with the WEKA tool revealed accuracies of 85.03% and 84.01% for Naive Bayes and Decision Tree, respectively. The article recommended employing a genetic algorithm in MATLAB to decrease the number of features before inputting the dataset into the WEKA tool for future research. In [20], Palaniappan and Awang introduced an Intelligent Heart Disease Prediction System (IHDPS) utilizing multiple classification techniques, including Decision Tree, Naive Bayes, and Neural Network. The system, implemented with the .NET framework, operates as a web-based platform. The dataset used comprises 909 records with fifteen attributes sourced from the Cleveland Heart Disease database. The model was established using the Data Mining Extension (DMX) query language. The outcomes indicated that Naive Bayes exhibited the highest efficiency with 86.53% correct predictions, closely followed by Neural Network with only a 1% difference.

In their work [21], Dangare and Apte created a model for heart disease prediction using the Cleveland database, which contains 303 records, and the Statlog database, comprising 270 records. Rather than utilizing solely the thirteen attributes inherent in the dataset, they introduced two additional attributes: obesity and smoking. The preprocessing of the dataset was carried out using the WEKA tool. The dataset was analyzed using classification techniques such as Decision Tree, Naive Bayes, and Artificial Neural Network (ANN). As per the outcomes, ANN achieved 100% accuracy, Decision Tree attained 99.62%, and Naive Bayes recorded 90.74%. This affirms that the Artificial Neural Network stands out as the most effective algorithm. In their work [22], Zriqat et al. created a proficient intelligent medical decision support system. The study involved a comparison of five classification algorithms: Naive Bayes, Decision Tree, Discriminant, Random Forest, and Support Vector Machine. MATLAB was employed for analysis on two datasets, the Cleveland Heart Disease and the Statlog Heart Disease. The findings indicated that Decision Tree exhibited the highest accuracy for both datasets, reaching 99.01% and 98.15% for the Cleveland and Statelog datasets, respectively.

In their study[23], Liu et al. introduced a hybrid model for heart disease diagnosis, utilizing the Statlog heart disease dataset sourced from the UCI repository.The MATLAB-developed model comprised two subsystems: feature selection and classification. The feature selection subsystem employs the Relief method to assess the weight of features, followed by the application of the Rough Set method (RFRS) to eliminate redundant features and enhance the model's accuracy. The classification subsystem utilized an Ensemble classifier with the C4.5 algorithm, which generates a Decision Tree as the base. The outcomes demonstrated a classification accuracy of 92.59%.

Ghumbre et al. conducted a comparison between Support Vector Machine and Radial Basis Function (RBF), a type of Artificial Neural Network (ANN). These algorithms were employed on a patient dataset from India, containing 214 records and 19 attributes, to predict the presence or absence of heart disease. The algorithms' performance was assessed by calculating the overall average through training and testing the dataset, along with

5-fold cross-validation and 10-fold cross-validation. The overall average performance resulted in accuracies of 86.42% for SVM and 80.81% for RBF. The findings indicated that SVM achieved superior accuracy. In their work [24], Masethe and Masethe utilized various algorithms, including J48, Naive Bayes, REPTREE, Simple Cart (Classification and Regression Tree), a type of Decision Tree, and Bayes Net for heart disease diagnosis. The dataset utilized in this study was acquired from South African physicians, encompassing eleven attributes: patient identification number (substituted with dummy values to safeguard patient privacy), gender, cardiogram, age, chest pain, blood pressure level, heart rate, cholesterol, smoking, alcohol consumption, and blood sugar level. The experiment utilized the WEKA tool, and the effectiveness of the constructed model was evaluated through a 10-fold cross-validation for performance assessment. The outcomes demonstrated accuracy rates of 99.0471% for J48, 99.0471% for REPTREE, 97.222% for Naive Bayes, 98.1481% for Bayes Net, and 99.0741% for the basic cart. These findings indicate that the simple cart surpassed the performance of the other models.

IV. Evaluation Various Machine Learning Approaches for Predicting Heart Disease

This section presents a tabulated comparison of the aforementioned research papers. The evaluation is conducted based on accuracy and is displayed in Table 2, comprising six elements as follows:

a) Authorship: This indicates the individual or individuals responsible for the paper and the corresponding reference number.
b) Classification Approach/es: This denotes the classification algorithm employed in the study, indicating whether a single algorithm, a comparative analysis, or a hybrid model was utilized.
c) Optimal Methodology Identified: This column is relevant solely to papers conducting comparisons among multiple algorithms. It signifies the most effective algorithm identified in the research, determined by the highest accuracy achieved.
d) Tool: This column displays the framework or programming language employed to construct the model. It represents the researcher's choice for pre-processing the input dataset, developing the predictive model, and conducting testing.
e) Dataset: This indicates the data set utilized as input for the classification algorithm.
f) Accuracy: This column signifies the precision of the outcomes from the proposed model. In cases where the paper includes a comparative analysis, it exclusively presents the accuracy of the best technique identified by the author.

Table 2 Evaluation of Heart Disease Approaches for Predicting Heart Disease

Authorship	Classification Approach/es	Optimal Methodology Identified	Tool Utilized	Dataset	Accuracy

Vembandasamy et al. [7]	NB	*n/a	WEKA	A diabetic research institute in Chennai	86.41%
Das et al. [9]	ANN Ensemble	n/a	SAS enterprise miner 5.2	Cleveland (UCI)	89.01%
Chen et al. [10]	ANN LVQ	n/a	C and C#	Cleveland (UCI)	80%
Dangre and Apte [11]	ANN	n/a	WEKA	Cleveland and Statlog (UCI)	Nearly 100%
Sabarinathan and Sugumaran [12]	DT	J48 with feature selection	Not mentioned	A dataset with 240 records for testing and 120 for training	85%
Shouman et al. [14]	KNN	n/a	Not mentioned	Cleveland (UCI)	97.4%
Wiharto et al. [15]	SVM	BT SVM	Not mentioned	Cleveland (UCI)	61.86%
Khateeb and Usman [3]	NB, KNN, DT and bagging technique	KNN	WEKA	Cleveland (UCI)	79.20%
Pouriyeh et al. [5]	NB, DT, MLP, KNN, SCRL, RBF, SVM, bagging, boosting and stacking	Boosting with SVM	Not mentioned	Cleveland (UCI)	84.81%
Waghulde and Patil [17]	ANN and Genetic Algorithm hybrid system	n/a	MATLAB	American Heart Association dataset	98%
Venkatalakshmi and Shivsankar [19]	NB and DT	NB	WEKA	UCI	85.03%
Palaniappan and Awang [20]	DT, NB and ANN	NB	DMX	Cleveland (UCI)	86.53%
Dangare and Apte [21]		ANN	WEKA	Cleveland and Statlog (UCI)	Nearly 100%
Zriqat et al. [22]	NB, DT, Discriminant, Random Forest, and SVM	DT	MATLAB	Cleveland and Statlog (UCI)	99.01% for Cleveland and 98.15% for Statlog
Liu et al. [23]	ReliefF and Rough Set (RFRS) for feature reduction,	n/a	MATLAB	Statlog (UCI)	92.59%

	Ensemble using C4.5 for classification				
Masethe and Masethe [24]	J48, NB, REPTREE, Simple Cart, and Bayes Net	Simple Cart	WEKA	South African dataset containing 11 attributes	99.07%

V. Conclusion:

This paper provides a survey of the existing literature on machine learning classification methods employed in the diagnosis of heart disease. Numerous papers, presenting various approaches using machine learning classification techniques, were reviewed and systematically categorized. The precision of the suggested models fluctuates based on factors such as the tool employed, the dataset utilized, the quantity of attributes and records within the dataset, the preprocessing techniques applied, and the classifier integrated into the model. This variability is influenced by factors such as whether the model is a hybrid, and whether it incorporates feature selection. Based on the findings in Table 2, it can be inferred that the researchers achieving the highest accuracy were Dangare and Apte, employing an Artificial Neural Network (ANN), the WEKA tool, and a fusion of the Cleveland and Statlog heart disease datasets.

In summary, constructing an accurate predictive model for heart disease necessitates the utilization of a dataset with an ample number of samples and accurate data. The dataset needs to undergo appropriate preprocessing, as it constitutes the most crucial step in preparing the data for utilization by the machine learning algorithm and achieving favorable results. Additionally, it is essential to employ an appropriate algorithm when constructing a prediction model. It is evident that Artificial Neural Network (ANN) consistently demonstrated strong performance across various models for heart disease prediction, along with Decision Tree (DT).

In conclusion, the application of machine learning in the diagnosis of heart disease is a significant domain, holding the potential to benefit both healthcare professionals and patients. The field is continuously evolving, and despite the abundant availability of patient data in medical facilities, there is limited publication of such data. Table 2 reveals that the majority of researchers sourced their datasets from the same origin, namely the UCI repository. Given that the dataset's quality significantly influences prediction accuracy, there should be greater encouragement for hospitals to release high-quality datasets (while maintaining patient privacy). This initiative would provide researchers with reliable sources to aid in model development and yield favorable results.

REFRENCES

[1] I. Kononenko, “Machine learning for medical diagnosis: history, state of the art and perspective,” *Artif. Intell. Med.*, vol. 23, no. 1, pp. 89–109, 2001, doi: https://doi.org/10.1016/S0933-3657(01)00077-X.

[2] J. Soni, U. Ansari, and D. Sharma, “Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers,” vol. 3, no. 6, pp. 2385–2392, 2011.

[3] N. Khateeb and M. Usman, “Efficient Heart Disease Prediction System using K-Nearest Neighbor Classification Technique,” *Proc. Int. Conf. Big Data Internet Thing*, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:4698456.

[4] H. Almarabeh and E. F., “A Study of Data Mining Techniques Accuracy for Healthcare,” *Int. J. Comput. Appl.*, vol. 168, no. 3, pp. 12–17, 2017, doi: 10.5120/ijca2017914338.

[5] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. R. Arabnia, and J. B. Gutierrez, “A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease,” *2017 IEEE Symp. Comput. Commun.*, pp. 204–207, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:12661480.

[6] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, “Big data preprocessing: methods and prospects,” *Big Data Anal.*, vol. 1, no. 1, p. 9, 2016, doi: 10.1186/s41044-016-0014-0.

[7] K. Vembandasamyp, R. R. Sasipriyap, and E. Deepap, “Heart Diseases Detection Using Naive Bayes Algorithm,” *IJISET-International J. Innov. Sci. Eng. Technol.*, vol. 2, no. 9, pp. 1–4, 2015, [Online]. Available: www.ijiset.com.

[8] D. S. Medhekar, M. P. Bote, and S. D. Deshmukh, “Heart Disease Prediction System using Naive Bayes,” 2013, [Online]. Available: https://api.semanticscholar.org/CorpusID:5800381.

[9] R. Das, I. Turkoglu, and A. Sengur, “Effective diagnosis of heart disease through neural networks ensembles,” *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7675–7680, 2009, doi: https://doi.org/10.1016/j.eswa.2008.09.013.

[10] A. H. Chen, S. Y. Huang, P. S. Hong, C. H. Cheng, and E. J. Lin, “HDPS: Heart disease prediction system,” in *2011 Computing in Cardiology*, 2011, pp. 557–560.

[11] M. Chaitrali, S. Dangare, M. Sulabha, and S. Apte, “a Data Mining Approach for Prediction of Heart Disease Using Neural Networks,” *Int. J. Comput. Eng. Technol. (IJCET)*, vol. 3, no. 3, pp. 30–40, 2012, [Online]. Available: http://ssrn.com/abstract=2175569.

[12] V.Sabarinathan and V.Sugumaran, “Prediction of Heart Disease Using Decision Tree,” *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 6, no. 3, pp. 530–532, 2016, [Online]. Available: https://www.researchgate.net/publication/339106269.

[13] S. B. Patel, “HeartDiseasePredictionUsingMachinelearningandDataMiningTechnique,” no. March, 2016, doi: 10.090592/IJCSC.2016.018.

[14] S. Mai, T. Turner, and R. Stocker, “Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients,” *Int. J. Inf. Educ. Technol.*, pp. 220–223, 2012, [Online]. Available: https://api.semanticscholar.org/CorpusID:13888525.

[15] Wiharto, H. K usnanto, and Herianto, “Performance Analysis of Multiclass Support Vector Machine Classification for Diagnosis of Coronary Heart Diseases,” *Int. J. Comput. Sci. Appl.*, vol. 5, no. 5, pp. 27–37, 2015, doi: 10.5121/ijcsa.2015.5503.

[16] S. U. Amin, K. Agarwal, and R. Beg, “Genetic neural network based data mining in prediction of heart disease using risk factors,” in *2013 IEEE Conference on Information & Communication Technologies*, 2013, pp. 1227–1231, doi: 10.1109/CICT.2013.6558288.

[17] N. P. Waghulde and N. P. Patil, “Genetic Neural Approach for Heart Disease Prediction,” 2014, [Online]. Available: https://api.semanticscholar.org/CorpusID:212587651.

[18] N. G. B. Amma, “Cardiovascular disease prediction system using genetic algorithm and neural network,” *2012 Int. Conf. Comput. Commun. Appl.*, pp. 1–5, 2012, [Online]. Available:

https://api.semanticscholar.org/CorpusID:15828888.

[19] B. Venkatalakshmi and M. V Shivsankar, “Heart Disease Diagnosis using Predictive Data Mining,” *Int. J. Innov. Res. Sci. Eng. Technol.*, vol. 3, no. 3, pp. 1873–1877, 2014.

[20] S. Palaniappan and R. Awang, “Intelligent heart disease prediction system using data mining techniques,” in *2008 IEEE/ACS International Conference on Computer Systems and Applications*, 2008, pp. 108–115, doi: 10.1109/AICCSA.2008.4493524.

[21] C. S.Dangare and S. S. Apte, “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques,” *Int. J. Comput. Appl.*, vol. 47, no. 10, pp. 44–48, 2012, doi: 10.5120/7228-0076.

[22] I. A. Zriqat, A. M. Altamimi, and M. Azzeh, “A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods,” no. April, 2017, [Online]. Available: http://arxiv.org/abs/1704.02799.

[23] X. Liu *et al.*, “A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method,” *Comput. Math. Methods Med.*, vol. 2017, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:14342000.

[24] H. D. Masethe and M. A. Masethe, “Prediction of Heart Disease using Classification Algorithms,” vol. II, pp. 22–24, 2014.