A Review of Heart Disease Classification Base on Machine Learning Algorithms

Mayaf Zakwa Hasan¹, Adnan Mohsin Abdulazeez²
mayaf.zakawa@dpu.edu.krd, adnan.mohsin@dpu.edu.krd
¹Technical College of Zakho, Duhok Polytechnic University, Kurdistan Region, Iraq
²Technical College of Engineering, Duhok Polytechnic University, Kurdistan Region, Iraq

Abstract
Heart disease is currently the leading cause of death. This problem is acute in developing countries. Predicting heart disease helps patients avoid it in its early stages and can also help medical practitioners find out the main causes. Machine learning has proven over time to play an important role in decision making and forecasting through massive data sets created by the healthcare sector. This review provides an overview of heart disease prediction using applied machine learning algorithms such as Naïve Bayes, Random Forest, Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression, and K-Nearest Neighbour (KNN). And these differences in the techniques are a reflection of many strategies for predicting heart disease. We present a synopsis of classification techniques that are primarily used in the predicted of heart disease. Additionally, we review several previous studies that conducted over the past four years, that used machine learning algorithms to predict cardiovascular.

Keywords
Heart Disease, Machine Learning (ML), Classification Algorithm.
A. Introduction

After the brain, the heart is one of the most important organs in the human body. The heart’s main job is to pump blood to every portion of the body [1]. If the heart isn’t beating correctly, other organs may cease working properly as well. Thus, maintaining the health of the heart and other organs becomes difficult [2], in any condition that may impair the heart’s ability to operate is called heart disease [1]. Heart disease claims the lives of 17.5 million people annually. These days [3].

The most prevalent illness in the world that poses a risk to human life also known as cardiovascular disease [4]. Heart disease encompasses a variety of heart-related disorders, such as blood vessel disease, heart failure, heart attack, arrhythmia, stroke, etc. [5]. Therefore, heart disease diagnosis or prognosis is necessary. Heart disease can be diagnosed using a variety of techniques. Although angiography is a widely used technique for identifying cardiac illness, doctors find it challenging to use it because to its high cost and intricate analysis [1], particularly in developing nations [4]. Reducing the risk of heart disease and stopping the rising death toll are largely dependent on early detection of heart disease and the use of suitable treatment [6]. As a result, non-invasive techniques for predicting cardiac disease have been developed; however, they are time-consuming and may produce inaccurate results [1]. The provision of high-quality services and accurate and efficient prediction is currently the main problem in the field of medical sciences [7]. We require an automated method in order to prevent these mistakes and produce better and quicker results. Researchers have discovered in recent years that machine learning algorithms are highly effective at analysing medical data sets [1].

Machine learning is an area that can help predict disease prognosis. Without special programming, machines can now learn thanks to machine learning [8]. Machine learning has applications in medicine that include diagnosing, detecting, and predicting diseases [7]. Predicting cardiovascular disease has become a new approach thanks to the use of machine learning. Using patient datasets, several machine learning methods are used to extract important information from the health club [9]. Many classifiers and algorithms, including K-nearest, Decision Tree, Random Forest, Support Vector Machine (SVM), Naïve Bayes, and others, can provide a solution to this situation thanks to advances in machine learning and artificial intelligence [8]. The field of machine learning is becoming more important for early diagnosis of diseases [6].

In medical applications, Classification is the most widely used machine learning technique in medical contexts since it is applicable to common real-world problems. Classification algorithms use training data to build a model first, which is then used on test data to produce predictions. Many categorization techniques have been used to diagnose diseases, with very positive results. These methods have the potential to produce quick results and lower diagnostic mistake rates [9].

The aim of this review is to provide a comprehensive overview of machine learning techniques in heart disease prediction. The structure of this paper is as follows. Section II discusses about Theoretical background ML and Algorithm of Heart Disease. Section III presents Related Work. Section IV Discussions. Finally, Section V presents Conclusion and Future Directions.
B. Research Method

1. Machine Learning

It is a subfield of artificial intelligence (AI) [10], [11], and is a large field of learning that deals with machines that mimic human capabilities [11]. Machine learning is frequently used in many different fields to address difficult problems that cannot be easily solved by computer-based methods [12]. Machine learning is one of the effective testing techniques that relies on machine learning training and testing [11]. It includes algorithms designed to perform various tasks including prediction, classification, and decision making. To facilitate the learning process, these algorithms require training data. After this learning phase, the algorithms create a model that serves as the output. This model is later tested and validated using unseen real-time test datasets. The final accuracy of the model is then compared with the actual values, proving the overall validity of the predicted results [13], [14].

Three different kinds of machine learning algorithms exist [11]:

![Classification of Machine Learning](image)

Figure 1 Classification of Machine Learning [11]

Techniques for machine learning are categorised as [14]:

a) Supervised Learning

A type of learning that occurs under supervision or with a mentor present is known as supervised learning. Here, a training dataset acts as the teacher for making predictions about a particular dataset. In other words, if testing data is used, there is always a matching training dataset that offers the direction needed to learn [11]. However, a drawback of supervised machine learning algorithms is that their ability to identify interesting patterns within the data is limited [15].

b) Unsupervised learning

is defined as self-contained learning that takes place in the absence of direction or instruction. The procedure is not being guided by an outside teacher. Unsupervised learning involves an algorithm that, given a dataset, analyses it on its own to find patterns and relationships. It then classifies and arranges newly introduced data inside the established patterns, based on these inferred associations [11]. Furthermore, it is possible to view unsupervised learning as more sovereign or agnostic [16].
1.1 Machine Learning Algorithms

This section offers a comprehensive review of the most commonly used machine learning algorithms for disease diagnosis [18].

1.1.1 Logical regression

Logistic regression is a supervised learning technique used to address binary classification issues [7], [8], [18]. Essentially, it works as a regression model that predicts the probability that an item or data input belongs to a particular category [8]. The logistic function is particularly valuable because it produces values between zero and one [4], [18], making it interpretable as a probability [4]. Logistic regression uses a sigmoidal function to model the data, as shown in Figure 3 [8]. Notably, logistic regression offers several key advantages including simplicity of implementation, computational efficiency, effectiveness based on training data, and ease of interpretation. In addition, input features do not require scaling [8].

![Logistic Function](image-url)

**Figure 3. Logistic Function [19]**
The Decision Tree (DT) serves as a supervised machine learning algorithm utilized for addressing regression and classification tasks [8], [13], accomplishing this by repeatedly segmenting data according to particular variables [8]. This model, resembling a tree structure, forms the basis of classification [7], [13]. Comprising root nodes, branches, and leaf nodes, the decision tree, analyses data by following the path from the root to the leaf node [8], [13]. Both numerical and categorical data are managed by the decision tree [7], [8]. Its primary objective is to construct a model capable of predicting a target variable by learning straightforward decision rules from training data [8]. The decision tree algorithm simplifies the interpretation of results, offering higher accuracy compared to alternative methods by organizing the dataset into a tree-like graph [7], as illustrated in Figure 4 [20]. However, the possible drawback is when working with large datasets and making decisions based on a single attribute at a time [7].

![Figure 4. Decision tree](image)

1.1.3 Support Vector Machine

The Support Vector Machine (SVM) stands as a machine learning algorithm employed in supervised learning scenarios [13], [21], [22], aimed at scrutinizing data and uncovering patterns in both classification and regression analyses [13], [22], [23]. Generally, SVM is considered when the data is classified as a two-class problem. The ideal hyper plane that isolates every data point from one class to the other class is what defines data in this strategy. The model is deemed better the greater the difference or edge between the two classes. Support vectors are the data points that are located near the margin’s edge [13]. The primary objective of a support vector machine is to identify the optimal highest-margin separating hyperplane between the two classes [7], [23]. Owing to its computational effectiveness, SVM has emerged as a proficient approach in addressing high-dimensional space problems, particularly with large datasets [23].
1.1.4 Naive Bayes

Naïve Bayes (NB) serves as a classification algorithm rooted in probability and statistics [8], [14], operates under a supervised learning paradigm [4], [14]. The basis of this approach is the Bayes theorem [14], [25]. And it has become a staple in machine learning applications owing to its straightforwardness in assigning equal importance to all features in the final decision-making process [8]. The reason for its extensive use is that it produces excellent results, outperforming even more complex categorization techniques [25]. The Naïve Bayes approach proves to be both versatile and apt across various domains. This technique boasts several advantages, such as its suitability for extensive datasets and its overall utility, making it applicable for both binary and multiclass classification challenges [8].
1.1.5 Random Forest

Random Forest is a supervised machine learning framework that may be used for both regression and classification applications [2],[26],[27]. It comprises a collection of trees [2],[28], with each tree constructed through repeated sampling of the dataset. Prediction is made by aggregating the outputs of all individual decision trees [2]. Increasing the number of trees is believed to enhance the accuracy and reliability of the model's predictions [2],[26]. Random Forest is particularly effective for large datasets and can also handle data with high dimensionality [2]. By considering the predictions of each tree, Random Forest produces a final output by selecting the class with the highest probability, determined through majority voting [27],[28]. Figure 7 provides an illustration of selecting procedure [28].

![Random Forest Diagram](image)

**Figure 6.** Random Forest prediction [28]

1.1.6 K-nearest neighbour Algorithm (KNN)

The KNN algorithm stands out as one of the simplest algorithms to grasp [4],[29],[30]. It is used for pattern recognition utilising a non-parametric approach for classification and inference [4]. In order to classify data points according to their proximity to one another, KNN measures the distance between each one [11]. Large datasets are especially well-suited for KNN classification, which requires more processing time during testing than training [4],[31]. However, it has strong computing powers [27]. KNN uses training datasets to forecast where K-nearest neighbours will be, and it uses Euclidean distance to measure how close the training dataset is to the target [29]. Nevertheless, a shortcoming in the KNN method is its susceptibility to local data structures [4].
1.1.7 Gradient Boosting

Gradient boosting is a machine learning technique that solves regression and classification problems [33], [34], by building a predictive model from a collection of weak predictive models [33]. An ensemble of decision trees is called gradient boosting. The approach relies on learning decision trees iteratively in order to minimise the loss function. Gradient boosting can handle nonlinearities and deal with categorical information because of the properties of decision trees [33]. The algorithm's strength is its capacity to train and enhance the classifiers iteratively, leading to an ensemble model that capitalises on each classifier's unique strengths and ultimately improves overall classification accuracy [34]. Boosting is a technique used to elevate underperforming models to a higher level of proficiency. In boosting, every subsequent tree is trained on a revised rendition of the initial dataset [33].

1.1.8 AdaBoost (Adaptive Boosting)

AdaBoost is an iterative process developed by to increase the performance of weak classifiers, sometimes known as base classifiers. Through repeated correction of these base classifiers' mistakes, AdaBoost improves the data's classification performance. This repeated refining helps to reduce variance and bias. The algorithm's efficacy stems from its ongoing iteration of classifiers, which leads to an ensemble model that optimises each classifier's distinct strengths and improves overall classification accuracy [34].

1.1.9 Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) are a component of neural networks within machine learning. ANNs operate in a manner akin to the human brain, resembling the structure and function of human neuron cells [1], [2]. They find applications across various domains of medicine [35] and serve as a non-linear statistical architecture for tackling complex problem-solving tasks [2]. This model comprises numerous interconnected elements, or neurons, collaborating to
execute a task. A neural network with a single layer is termed a perceptron, producing a singular output [1], as depicted in Fig. 10.

Figure 8. Artificial Neural Network [36]

C. Literature Review

In [37], the researchers have focused on how machine learning (ML) can be used to predict and understand heart disease symptoms. They revealed several notable factors, including heart rate, cholesterol, chest pain, features associated with ST depression, and vascular characteristics of the heart. According to their study, combining chi-square and PCA enhances performance for most classifiers, but applying PCA directly from raw data leads to less-than-ideal results.

In [38], the researchers have proposed a systematic approach to machine learning-based heart disease detection. To increase accuracy and efficiency, they used several feature selection strategies and classification algorithms. Cross-validation was used to evaluate the performance and a new feature selection approach was presented. They discovered that the proposed technical results are successful and applicable, superior to previous methods and have the potential for use in healthcare applications.

In [39], the authors proposed using data mining and machine learning techniques to predict heart disease using a variety of algorithms. Emphasis was placed on the importance of early detection and the function of machine learning in medical diagnosis. Positive results from one model suggested that it may be useful in predicting heart disease. She indicated that alternative models could be studied in further research to increase the accuracy of diagnosis.

In [40], the study focused on applying machine learning algorithms in classifying cardiovascular data sets for the purpose of disease prediction. They emphasized how important the dimensions of the data set are to the performance of the algorithm and plan to improve disease prediction models in the future by reducing their dimensions. Their main goal was to provide information on early diagnosis and management of cardiovascular diseases.
In [41], the authors proposed using several feature selection strategies and classification algorithms to examine how well machine learning models diagnose heart conditions. Principal components analysis and chi-square testing were used to identify separate feature sets from the four heart disease datasets analyzed. They found that the best accuracy is achieved when combining the BayesNet method and chi-square feature selection. They emphasized the importance of feature selection in enhancing heart disease prediction algorithms.

In [42], the researchers focused on highlighting the importance of data analytics, especially when it comes to applying machine learning algorithms to predict heart disease. They used heart disease datasets to evaluate different supervised learning techniques. They recommend future investigations using larger data sets and other machine learning methods to create more accurate real-time prediction models for heart disease.

In [14], this study explored different data mining techniques for predicting heart disease, using supervised learning algorithms such as Naïve Bayes, decision tree, K-nearest neighbor, and random forest. Using a dataset from the Cleveland Database, the study evaluated 14 features to evaluate the performance of the algorithm. K-nearest neighbor showed the highest accuracy. They suggest that future research should consider additional data mining techniques to enhance predictive accuracy for early detection of heart disease.

In [43], the study proposes a clinical support system that predicts heart disease using machine learning algorithms, to help doctors make more accurate diagnoses and decisions. Naïve Bayes emerged as the best-performing algorithm, showing notable accuracy in both split and cross-training tests. It has been recognized that it can be applied to smaller data sets. The results highlighted the potential of the predictive model to enhance heart disease risk prediction, thereby improving clinical diagnosis and treatment decisions.

In [2], this research aims to help medical professionals diagnose and treat patients by examining the importance of using machine learning methods to predict heart disease, given the vast amount of medical data available. Through comparative analysis, random forest emerged as the most effective algorithm for predicting heart disease, as determined by evaluation and accuracy metrics. They suggested that future studies should focus on implementing deep learning algorithms to enhance predictive capabilities.

In [44], the researchers worked on automatic detection of cardiovascular diseases using different machine learning algorithms. They found that the Gradient Boosting Classifier performed best across all features, and Random Forest was the best classifier. Which helped doctors diagnose and make decisions faster and more efficiently in cardiovascular diseases.

In [45] discussed the study on the effectiveness of machine learning algorithms in predicting heart disease in order to overcome difficulties in clinical decision-making and diagnostic accuracy. They found the possibilities of accurate disease prediction, using supervised algorithms such as KNN, DT and RF. They found key predictors using feature importance ratings, demonstrating the value of machine learning in enhancing patient care and outcomes.
In [46], the authors addressed the issue of major cardiac anomalies. They proposed a new method for real-time feature extraction (less than 30 milliseconds) and ECG type prediction using machine learning and digital ECG data. They implemented the XGBoost algorithm. They created a standard for detecting heart abnormalities. It has made great progress and shown excellent performance in hospitals and countries.

In [47], the study worked on using deep learning and machine learning methods to predict heart disease. Using the 1988 data set, in order to improve accuracy, the researchers compared several classifiers and methodologies. He highlighted pre-processing, normalization, and outlier data identification. Which led to the development of computational intelligence and neuroscience.

In [48], the researchers have proposed a variety of feature selection techniques to examine how well classification models predict heart disease. The primary goal was to evaluate the effects of 10 feature selection strategies and six classification algorithms on the accuracy of heart disease prediction. They highlighted how important feature selection is in improving the success of prediction algorithms in diagnosing heart disease.

In [49], the study addressed the topic of machine learning (ML) in heart disease diagnosis by using machine learning algorithms, including supervised techniques such as Random Forest and K-Nearest Neighbor (KNN), to classify people based on characteristics such as age, cholesterol and weight. Health status. And chest pain. The results highlight the effectiveness of machine learning in diagnosing cardiovascular disorders.

In [50], the researchers presented a comprehensive model designed to accurately predict heart disease. By using LASSO and Relief methods to select features, perform pre-processing and collect data in an efficient manner. Moreover, they developed state-of-the-art hybrid classifiers that combine traditional classifiers with bagging and boosting techniques, such as DTBM, RFBM, KNNBM, ABBM, and GBBM. They evaluated typical performance indicators using machine learning methods.

In [51], the authors presented the development of a model that uses three machine learning classification techniques to identify cardiovascular disease. Their model has proven to be highly robust and accurate, correctly identify people at risk of heart disease, and reducing pressure on healthcare systems. It has reduced expenses while improving medical treatment. Their study, conducted in Pynb format, provided insight into predicting patient heart disease.

In [52], the authors proposed developing a diagnostic framework using several machine learning techniques to classify heart diseases. Its goal is to improve prediction accuracy by optimizing model parameters for effective detection of heart disease. They pointed out the possibility of an effective health care monitoring system for early detection of heart disease. Data mining classification techniques have accelerated the learning process, simplified models, and improved prediction accuracy, especially when combined with optimization methods.

In [53], the study proposes a hybrid decision support system that can identify heart problems early and is particularly useful in settings where access to specialized health care is scarce. Among the pre-processing techniques, mixed
feature selection and multivariate imputation. Used a variety of classifiers, their system aimed to reduce delays in diagnosing heart disease while improving accuracy compared to current techniques.

In [54], the research discussed the predictive accuracy of eleven machine learning classifiers for heart disease risk. To select the best features, they used three feature evaluators, and adjusted hyperparameters to improve accuracy. Prediction accuracy has been greatly increased by this approach. The researchers have advised future research on different feature selection algorithms and methods. He also intended to merge several data sets in order to collect more information and conduct more research in order to increase the accuracy of predictions.

In [55], the researchers worked to creating a machine learning model that used the DiScRi dataset to predict the incidence of diabetes and cardiovascular disease. They identified algorithm risk variables that affect the chance of developing the condition, such as family history, HbA1c, and HDL. The model provided the possibility of early identification and treatment in prevention.

In [56], the researchers used a variety of algorithms combined with sequential feature selection, using K-fold cross-validation for validation purposes. Their analysis, using datasets from a variety of sources, showed that Random Forest and Decision Tree classifiers achieved remarkable levels of accuracy. Their future efforts seek to expand the model’s reach and utility, and they highlighted advances in feature selection methods and considering incorporating a random forest classifier to improve predictive performance.

In [57], the authors describe how to prediction of cardiovascular disease using logistic regression. They stressed the importance of early detection to receive timely treatment. They prepared the data after cleaning, filling in missing values, and selecting strongly associated characteristics. They have used the classification process for many courses and test sections. They recommended further research using multiple datasets.

In [58], the researchers focused on applying machine learning algorithms to predict heart disease. In order to solve the problem of early diagnosis and lower mortality rates. They used a variety of prediction algorithms and collected data sets. The goal of their study was to improve the accuracy of predicting the rate of heart disease by examining a large amount of medical data. They have advanced the field of medicine by using data science to enhance the prediction of heart disease.

In [59], the study proposed using machine learning algorithms to predict heart disease, to improving accuracy and reliability in diagnosis. They used classifiers and evaluated metrics such as sensitivity and specificity. Preprocessing and hyperparameter tuning improved prediction accuracy, they highlighted the potential of machine learning in cardiovascular healthcare.

In [60], the researchers discussed how to using the capabilities of the diagnostic system and applying heuristic algorithms to classify heart disorders. They relied on a reduced set of features and revealed several different classifications to predict the presence of heart disorders. They found that the results achieved the highest accuracy among the algorithms tested, which
confirmed their usefulness in increasing prediction accuracy for diagnosing heart diseases.

In [61], the authors have proposed using machine learning techniques to improve the accuracy of heart disease prediction. Using GridSearchCV, a five-fold mutual improvement of hyperparameters was achieved, introducing innovation. Real-time heart disease prediction has been made possible through the potential integration of the model with useful applications such as online platforms or mobile devices, which have leveraged patients' electronic health records to improve accuracy.

In [62], the researchers presented the development of a machine learning model for predicting CVD risk. Different machine learning models were compared, which are useful for early diagnosis and more effective treatment outcomes. Their goal was to address the problem of cardiovascular diseases (CVDs), and they also emphasized how machine learning algorithms can improve the effectiveness of identifying and treating diseases.

In [63], the research discussed machine learning algorithms to predict and diagnose cardiovascular problems early. The aim of their research was to improve the diagnostic and therapeutic skills of medical professionals. The performance of several algorithms has been evaluated and they used in global healthcare. To increase accuracy, they advised further research into machine learning methods and patient data.

In [6], the authors worked on a supervised machine learning approach called Classification and Regression Tree (CART) algorithm for heart disease. They demonstrated that the model is reliable in predicting heart disease. Furthermore, the model's decision criteria simplify its implementation in clinical settings and eliminate the need for additional knowledge. The CART model has made it easier to identify and manage heart disease, and is a vital tool for patients and healthcare professionals alike.

<table>
<thead>
<tr>
<th>Authors and year of pub</th>
<th>Dataset</th>
<th>Method</th>
<th>Pros</th>
<th>Limit</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Escamilla et al 2020 [37]</td>
<td>Cleveland-Hungarian</td>
<td>DT, Gradient-Boosted Tree, LR, Multilayer Perceptron, Naïve Bayes, and RF</td>
<td>Improved prediction accuracy, Dimensionality reduction, Relevance of features</td>
<td>Small sample size, Limited extension to other diseases</td>
<td>98.7% Cleveland, 99.0% for Hungarian, 99.4% Cleveland-Hungarian</td>
</tr>
<tr>
<td>Li et al 2020 [38]</td>
<td>Cleveland Heart Disease dataset</td>
<td>Sequential Backward Selection, KNN, ANN, SVM, Naïve Bayes, DT</td>
<td>High accuracy, used various algorithms, Information feature selection algorithm.</td>
<td>Research gap in improving prediction accuracy and reducing calculation time.</td>
<td>80.4% to 92.37%</td>
</tr>
<tr>
<td>Obiri and Sarku 2020 [39]</td>
<td>Cleveland Heart Diseases</td>
<td>Linear Regression, Decision Tree, Gaussian Naïve Bayes</td>
<td>Improved accuracy in diagnosis and the potential for reducing the number of tests needed for diagnosis.</td>
<td>82.75% 79.31% 76%</td>
<td></td>
</tr>
<tr>
<td>Princy et al 2020 [40]</td>
<td>cardiovascular disease dataset from Kaggle</td>
<td>Decision Tree, Logistic Regression, Random Forest, Support Vector Machine with RBF Kernel, Naïve Bayes, and K-Nearest Neighbors</td>
<td>Comparison of pre and post dimensionality reduction accuracy of algorithms, and the implementation of various classification techniques for predicting.</td>
<td>The model requires larger dataset testing, interprets Neural Network processes more complexly, and dataset dimensionality significantly impacts algorithm performance. 0.73</td>
<td></td>
</tr>
<tr>
<td>Spencer et al 2020 [41]</td>
<td>Cleveland, Long-Beach-VA-Dataset, Hungarian, Switzerland</td>
<td>BayesNet, LR, Stochastic Gradient Descent, KNN IBK with K, Adaboost M1 with Decision Stump, Adaboost M1 with Logistic, RF</td>
<td>Evaluate machine learning models, compare different feature selection methods to enhance accuracy and identify influential features.</td>
<td>A few machines learning and feature selection methods. 80% to 95.2% 85%</td>
<td></td>
</tr>
<tr>
<td>Sujatha and Mahalakshmi 2020 [42]</td>
<td>Heart disease dataset obtained from the Kaggle</td>
<td>Decision Tree, Naïve Bayes, Random Forest, Support Vector Machine, K-Nearest Neighbor, and Logistic Regression.</td>
<td>High accuracy was achieved using the Rapid Miner tool. Determine the feature selection technique.</td>
<td>Limited discussion of the specific difficulties encountered and no comparison with other studies. 83.52%</td>
<td></td>
</tr>
<tr>
<td>Shah et al 2020 [14]</td>
<td>UCI</td>
<td>Naïve Bayes Classifier, Decision Tree, K-Nearest Neighbour, Random Forest Algorithm</td>
<td>High accuracy, data mining techniques complex medical insights into attributes.</td>
<td>Need to additional data mining techniques, limited discussion, need for more advanced models. 83% 80.26% 90.70% 86.84%</td>
<td></td>
</tr>
<tr>
<td>EL Hamdaoui et al 2020 [43]</td>
<td>Cleveland heart disease dataset</td>
<td>Naïve Bayes, K-Nearest Neighbour, Support Vector Machine, Random Forest, and Decision Tree</td>
<td>Accurate Prediction, Clinical Decision Support, Model Validation</td>
<td>Decrease in accuracy, dataset Size small, overfitting by using cross-validation. 82.17% 76.56% 79.20% 69.30% 75.57%</td>
<td></td>
</tr>
<tr>
<td>Katarya and Meena 2020 [2]</td>
<td>UCI</td>
<td>Logistic Regression, Naïve Bayes, SVM, KNN, Decision Tree, Random Forest, Artificial Neural Network, Deep Neural Network, Multilayer Perceptron</td>
<td>Thorough evaluation algorithms, providing insights into their effectiveness and enabling comparison of these algorithms.</td>
<td>Lack a more thorough examination of the computational resources needed to put the methods into practice. 95.60%</td>
<td></td>
</tr>
<tr>
<td>Alawawi and Alsuwat 2021 [44]</td>
<td>Cardiovascular Disease heart diseases</td>
<td>RF, SVM, DT Voting, LR, KNN, Naïve Bayes</td>
<td>Comprehensive Approach High Accuracy Utilization of Diverse Datasets Potential Clinical Impact</td>
<td>Limited Dataset Size Feature Engineering Challenges Comparison with Existing Methods 94% 65% 89% 67% 89% 71% 73%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Authors</th>
<th>Year</th>
<th>Dataset</th>
<th>Algorithms</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bertsimas et al 2021</td>
<td>[46]</td>
<td>(The 2017 Challenge dataset from Physionet), (Tianchi Hefei High-Tech Cup ECG), (University &amp; Shaoxing People’s Hospital)</td>
<td>XGBoost</td>
<td>The models outperformed original papers in predictive performance, enabling real-time detection of heart anomalies from ECG signals with predictions in under 30 milliseconds.</td>
</tr>
<tr>
<td>Bharti et al 2021</td>
<td>[47]</td>
<td>Cleveland, Hungary, Switzerland, Long Beach V.</td>
<td>LR, KNN, Decision Tree, RF, SVM, XGBoost</td>
<td>The study’s proprietary datasets, random data split between training and testing sets, and lack of time complexity information for cross-dataset models limit objective comparisons</td>
</tr>
<tr>
<td>Dissanayake &amp; Johar 2021</td>
<td>[48]</td>
<td>Hungarian, Cleveland, and Hungarian-Cleveland</td>
<td>decision tree, random forest, support vector machine, K-nearest neighbor, logistic regression, and Gaussian naive Bayes.</td>
<td>various algorithms, Comparison of different classifiers and technique, comprehensive dataset not capture the most current trends or data, focuses on a specific set of attributes and features, 88.52%</td>
</tr>
<tr>
<td>Garg et al 2021</td>
<td>[49]</td>
<td>dataset used from a website Kaggle</td>
<td>K-Nearest Neighbor Random Forest</td>
<td>Effective Prediction, Healthcare Improvement, Balanced Data specific algorithms, dataset dependency, scope of analysis 86.88%</td>
</tr>
<tr>
<td>Ghosh et al 2021</td>
<td>[50]</td>
<td>Cleveland, Hungary, Switzerland, VA Long Beach, and Statlog</td>
<td>Decision Tree, Random Forest, K-Nearest Neighbors, AdaBoost, and Gradient Boosting</td>
<td>improved prediction accuracy, providing a tightly correlated feature set for use with ML algorithms, need for larger datasets, Dependency on specific feature selection techniques, may limit the generalizability of the model, 99.05%</td>
</tr>
<tr>
<td>Jindal et al 2021</td>
<td>[51]</td>
<td>UCI</td>
<td>K Nearest Neighbors Logistic Regression Random Forest Classifier</td>
<td>ability to predict and classify patients with heart disease accurately, and the reduction in medical costs, need for further research to enhance prediction accuracy in Intelligent Heart Disease Prediction System (HDPS) model, 87.5%</td>
</tr>
<tr>
<td>Authors</td>
<td>Year</td>
<td>Datasets/Techniques</td>
<td>Performance/Conclusion</td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>------</td>
<td>-------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Patro et al.</td>
<td>2021</td>
<td>Cleveland, datasets from Beijing and China's environmental monitoring stations, and the Heart Statlog</td>
<td>Salp Swarm, Salp Swarm Optimized Neural Network, Bayesian Optimization, KNN, Naïve Bayes, SVM, RapidMiner, Particle Swarm Optimization, Ant Colony Optimization, and Fast Correlation-Based Feature Selection</td>
<td>improved accuracy, convergence, speed, and reliability. less control over multimodal strategies</td>
</tr>
<tr>
<td>Rani et al.</td>
<td>2021</td>
<td>University of California, Irvine (UCI)</td>
<td>hybrid system, support vector machine, naïve bayes, logistic regression, random forest, and adaboost classifiers</td>
<td>better accuracy compared to existing systems. It can assist doctors in making quick decisions. only diagnose the presence or absence of heart disease, not the severity.</td>
</tr>
<tr>
<td>Reddy et al.</td>
<td>2021</td>
<td>Cleveland</td>
<td>Naïve Bayes, LR, Sequential Minimal Optimization, Instance-Based Classifier, AdaBoostM1 with (DS), AdaBoostM1 with LR, Bagging with REP Tree, Bagging with LR, Jrip</td>
<td>Improved Forecasting Performance: hyperparameter tweaking and attribute selection strategies. Small Dataset Size, Limited Feature Selection Methods</td>
</tr>
<tr>
<td>Abdalrade et al.</td>
<td>2022</td>
<td>DiScRi</td>
<td>Logistic Regression, Evimp Functions, Generalised Cross-Validation, correlation matrix, Classification and Regression Trees</td>
<td>This prediction model can help with early diagnosis, preventative management, and lessening the cost of DM and CVD-related healthcare.</td>
</tr>
<tr>
<td>AHMAD et al.</td>
<td>2022</td>
<td>Heart Statlog, Cleveland Hungary, Heart Disease Clinical Record Data Set 2020</td>
<td>Linear Discriminant Analysis, Random Forest, Gradient Boosting Classifier, Decision Tree, Support Vector Machine, K-Nearest Neighbours</td>
<td>comparative study of ML techniques, introduces a hybrid approach, feature selection methods. Lacks detailed data set information and discussions of model biases. Further validation of the data sets, interpretability of the model and its clinical relevance.</td>
</tr>
<tr>
<td>Ambrish et al.</td>
<td>2022</td>
<td>UCI</td>
<td>LR, SVM, RF, Stochastic Gradient Boosting, J48, and Multilayer Perceptron.</td>
<td>high accuracy achieved compared to previous research work. limited dataset 87.10%</td>
</tr>
<tr>
<td>Bora et al.</td>
<td>2022</td>
<td>UCI heart disease, heart disease</td>
<td>Logistic Regression, SVM, RF, Naïve Bayes, KNN, Gradient Boost, XGBoost</td>
<td>comprehensive analysis and, potential for early diagnosis. limitations involve data quality dependency and scalability challenges. 73% to 96%</td>
</tr>
<tr>
<td>Year</td>
<td>Authors</td>
<td>Dataset</td>
<td>Features</td>
<td>Methodology</td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>---------</td>
<td>----------</td>
<td>-------------</td>
</tr>
<tr>
<td>2022</td>
<td>Saboor et al.</td>
<td>UCI, Z-Alizadeh Sanitation Heart, Hungarian and VA, Cleveland, Hungary, Switzerland</td>
<td>KNN, Neural Networks, SVM, Genetic Algorithms, J48 (Decision Trees), RF, Naïve Bayes</td>
<td>Dataset standardization and refinement, high accuracy</td>
</tr>
<tr>
<td>2022</td>
<td>Ullah et al.</td>
<td>Cleveland</td>
<td>DT, neural network, SVM, KNN, LR.</td>
<td>High accuracy, proving prediction accuracy through cooperative techniques</td>
</tr>
<tr>
<td>2023</td>
<td>Chandrashekhar and Peddakrishna</td>
<td>UC Irvine ML Repository - Cleveland IEEE Dataport heart disease</td>
<td>RF, KNN, LR, Naïve Bayes, Gradient Boosting, AdaBoost</td>
<td>Enhanced Prediction Accuracy, novel approaches such as GridsearchCV, Comparative Analysis</td>
</tr>
<tr>
<td>2023</td>
<td>Dalal et al.</td>
<td>UCI</td>
<td>QUEST, random forest, neural network, Bayesian network</td>
<td>Use of advanced machine learning approaches, and early identification and therapy for at-risk patients</td>
</tr>
<tr>
<td>2023</td>
<td>Khan et al.</td>
<td>Khyber Teaching Hospital and Lady Reading Hospital in Pakistan</td>
<td>DT, RF, LR, Naïve Bayes, SVM</td>
<td>Potential for effective prediction and diagnosis in hospital settings</td>
</tr>
<tr>
<td>2023</td>
<td>Ozcan and Peker</td>
<td>heart disease dataset</td>
<td>Classification and Regression Tree</td>
<td>Provides a high accuracy rate, the model is easily integrated into clinical decision support systems</td>
</tr>
</tbody>
</table>

### D. Result and Discussion

The table gives a summing up view of articles concentrating on different years from 2020 to 2023. They are based on multiple ways of predicting heart diseases. The findings present valuable insights, serving as a foundation for potential future research, showing the accuracy of the algorithms such as support vector machines and random forests. Logistic regression, decision trees and neural networks. It is interesting to note that in analysis of these very complex medical data sets, the capability to learn by machines is also emphasized. First, we will involve datasets like the Cleveland Heart Disease Dataset and those provided by sources like Kaggle, UCI, and hospitals as inputs. Accuracy rates reported in these studies show wide variation, ranging from approximately 68% to 100%. This diversity illuminated how algorithm or dataset
qualities vary and also affects the predictive performance. Increased prediction accuracy, reduced dimensionality and extraction of essential attributes are some of the advantages reported in researches. These advantages highlight how machine learning can be used to increase the effectiveness and efficiency of heart disease prediction algorithms. Although research demonstrates that there are some positive results, the limitations, which include small sample sizes, the inability to generalize to other conditions, and inadequate increase in predictive accuracy as well as taken time, clearly suggest further research work.

Second, in later research, it should be investigated how the details of data collection and model performance relate to the results. Routine medical practice will be extremely shaped by the data from these tests. Machine learning-based models for prediction can substantially decrease the overall costs of cardiovascular disease and intensify early identification and early intervention, thereby lowering personal and population outcomes. Despite that, but for a careful consideration of these factors, the models could hardly be used to measure efficiency and outcome of such care.

E. Conclusion and Future Directions

In conclusion, the main findings emerged that machine learning techniques were used across different datasets, methodologies and algorithms to achieve high accuracy rates in predicting heart disease. Even despite the notable progress in predictive accuracy and identification of relevant features, there are still many hurdles to overcome, such as small sample sizes, limited application to other diseases, and the need for further research to increase prediction accuracy and shorten calculation times. Moreover, in future research endeavors, issues such as data completeness, model performance evaluation, and scalability limitations must be carefully considered. Despite these obstacles, the results of these studies demonstrate the potential of machine learning for early diagnosis of cardiovascular disease, preventive care, and cost-saving measures. The main goals of future research efforts should focus on addressing these obstacles and leveraging the advantages of machine learning to improve cardiovascular disease management and prediction.

Subsequent research efforts ought to focus on addressing the identified challenges and using machine learning’s potential in predicting and managing cardiovascular illness. This means exploring new techniques, improving the caliber and quantity of datasets, enhancing the interpretability of models, and carrying out exhaustive validation tests in a range of clinical settings.

F. References


