Acta Scientiarum
http://www.periodicos.uem.br/ojs/
ISSN on-line: 1807-863X
Doi: 10.4025/actascibiolsci.v47i1.73238

Performance evaluation of machine learning techniques for heart disease prediction: An overview

Dhanashri Shankar Karande* and **Shailendrakumar Mahadeo Mukane**

Department of Electronics & Telecommunication Engineering, Shivnagar Vidya Prasarak Mandal's College, COE, Malegaon (Bk)-Baramati, SPPU, Pune, Maharashtra, India. *Author for correspondence. E-mail: dskraut@gmail.com

ABSTRACT. One of the most common and serious diseases is heart disease, as it is one of the major causes of death globally. Heart disorders come in many forms, including arrhythmia, congenital heart disease, and atherosclerosis. Patients with heart disease experience a variety of symptoms, such as dizziness, chest discomfort, and excessive perspiration. Heart disease is primarily caused by smoking, high blood pressure, diabetes, obesity, and other risk factors. Developing an affordable and non-invasive approach to predicting heart disease is necessary. Creating a system that accurately predicts heart disease with minimal errors is essential. Consequently, machine learning is vital for predicting the risk of future cardiopathy by analysing the patient's health conditions and past medical history to decrease the possibility of mortality from heart disease. Machine learning (ML) has rapidly advanced in recent years, and its application in medical sciences can revolutionize how complex diagnostic and prognostic evaluations are conducted at the individual patient level. ML helps predict the risk of developing heart disease based on the patient's current and historical medical conditions to decrease the chances of death due to heart disease. ML techniques, including Random Forest, ANN, Linear Regression (LR), Logistic Regression (LR), K-Nearest Neighbor (KNN), Naive Bayes (NB), Support Vector Machine (SVM), Gradient Boosting, and Decision Tree (DT), are utilized to create the machine learning model. This paper presents an overview of the Heart Disease Prediction system using Machine Learning techniques. A detailed tabular comparison of the reviewed papers is also included.

Keywords: decision tree; heart disease; linear regression; machine learning, and random forest.

Received on August 2, 2024
Accepted on May 9, 2025

Introduction

One of the leading causes of death in the modern world is heart disease. According to the World Health Organization [WHO] (2020), heart-related diseases account for 17.9 million annual mortalities worldwide. The heart pumps blood from the lungs to the lower body, and if blood circulation is compromised, other organs, like the brain, will suffer. If the heart stops pumping, death could occur in a matter of minutes. The heart's ability to pump blood is crucial for survival, and several factors—such as high blood pressure, high cholesterol, family history, smoking, an unhealthy diet, obesity, and inactivity can increase the risk of heart disease (Latha & Jeeva, 2019). Due to its high fatality rate globally, heart disease has become a significant health concern for many people.

Heart disorders can occur in a variety of forms:

Arrhythmia: The term "arrhythmia" refers to an abnormal heartbeat.

Atherosclerosis refers to the hardening of the arteries.

Cardiomyopathy: This disease leads to stiffer or weaker heart muscles.

Angina refers to chest pain caused by insufficient blood flow to the heart muscles.

Congenital heart defects are anomalies of the heart that are present at birth.

Coronary artery disease (CAD) is a disorder caused by the buildup of plaque in the arteries that supply blood to the heart. It is also referred to as ischemic heart disease.

Bacteria, viruses, or parasites could bring on heart infections.

The crucial behavioral risk factors for heart failure and stroke include poor nutrition, inactivity, smoking, and alcohol misuse. Patients with heart disease experience a variety of symptoms such as chest discomfort, dizziness, and profuse perspiration. People may develop high blood pressure, high blood sugar, high blood

lipids, overweight, and obesity due to these behavioral risk factors. These "intermediate risk variables" can be assessed in primary care settings and indicate an increased risk of outcomes, including heart attack, heart failure, and stroke (Diwakar et al., 2020).

In India, heart disease claims the lives of approximately 17 million people annually, and by 2030, it's predicted that this number will rise to 23 million. Therefore, a non-invasive, less expensive technique for predicting heart disease is needed. This is why researchers have become interested in predicting heart disease and have developed a model using various machine learning algorithms. Some of these algorithms achieved better results than others. Many used the well-known UCI heart disease dataset to train and test their classifiers, while others used data obtained from other hospitals available to them. (Ali et al., 2019)

This review paper provides an overview of the ML classification techniques used in diagnosing heart disease and how previous researchers implemented them. It shows how vital machine learning is in healthcare and how it can make accurate predictions and help healthcare professionals.

The rest of the paper is organized as follows: The Materials & Methods section outlines the methodology, encompassing machine learning and performance evaluation metrics. The Related Work section includes a literature review of current proposed research in this area. The table section offers a tabular comparison between the classification techniques based on their accuracy. The final section of the paper presents the discussion and conclusion.

Material and methods

Machine learning

Recent technological advancements in ML may significantly alter the usage of comprehensive diagnostic and prognostic assessments at the individual patient level. The most common types are supervised, unsupervised, and reinforcement learning (Shah et al., 2020a). ML is classified as follows (Figure 1):

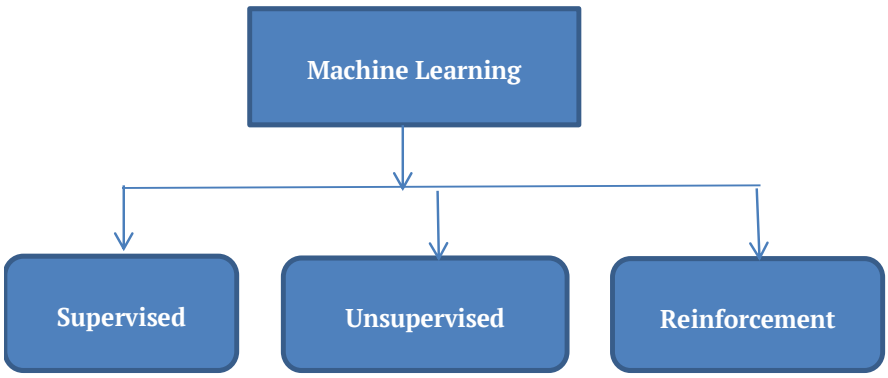

Figure 1. Classification of machine learning.

Supervised learning

Given a collection of observations, this algorithm models the relationship between independent variables and a label. The label of new data is then predicted using the model, based on the features. Depending on the characteristics of the target variable, this could be a classification task or a regression task.

Unsupervised learning

This method identifies organizations in unlabeled data. The incoming data lacks tags, so association rule learning and clustering techniques are employed to infer the relationships that exist within the data. The core principle of reinforcement learning is the action-reward system. An agent learns to accomplish a task by continuously evaluating the benefits of its actions.

Reinforcement learning

Reinforcement learning emphasizes the need to respond to the environment in order to optimize anticipated rewards. Data serve as feedback for the model. Online planning is a primary focus of reinforcement learning, which requires a balance between exploitation and exploration.

Performance evaluation metrics

Researchers utilize the metrics below to assess prediction models and demonstrate their performance results. We provide:

i) Accuracy: This metric indicates the percentage of accurate results.

ii) Precision: This metric indicates the relevance of the result.

iii) Recall or Sensitivity: Assesses the relevant results that are returned.

iv) F-Measure: It combines precision and recall.

v) Receiver Operating Characteristic (ROC): This graph visualizes the classifier's performance by displaying both correctly and incorrectly classified cases.

The most widely used performance evaluation metric is accuracy, which appears in all the research papers discussed in our article. Therefore, this overview article focuses on categorizing, comparing, and reviewing previous work based on accuracy.

Related work

This section contains various researchers' related work using machine learning techniques to build a model.

Gavhane et al. (2018) proposed a ML algorithm to predict the vulnerability to heart disease based on basic symptoms such as age, sex, and pulse rate. The researchers aim to develop an application that can predict the severity of heart disease using these symptoms. The machine-learning algorithm, neural networks, has proven to be the most accurate and reliable in predicting the chances of heart disease. The researchers proposed collecting relevant data related to their field of study, training this data using the proposed machine-learning algorithm, and predicting the likelihood of a patient contracting heart disease. The system will be user-friendly, allowing for regular monitoring of the patient. The study focuses on parameters with a significant risk percentage concerning CAD, such as age, sex, blood pressure, and heart rate.

Ambrish et al. (2022) discussed predicting heart disease using logistic regression, a ML algorithm. The dataset used for the analysis contains 13 features and 303 records. The paper focuses on data preprocessing, feature selection, and splitting the dataset into training and testing sets. Logistic regression is chosen as the classification algorithm and achieves an accuracy of 87.10%.

Dwivedi (2016) discussed the use of various ML techniques for predicting the presence or absence of heart disease. It evaluates the performance of six machine learning methods - Artificial Neural Network (ANN), Support Vector Machine (SVM), Logistic Regression (LR), k-nearest neighbour (kNN), classification tree, and Naive Bayes (NB) - on a heart disease dataset. The results show that LR achieved the highest classification accuracy. The document also presents the performance of these methods in terms of other metrics such as precision, negative predictive value, and F1-measure.

Latha and Jeeva (2019) analyzed the accuracy of heart disease predictions using an ensemble of classifiers. The Cleveland heart dataset from the UCI machine learning repository was utilized for training and testing. The ensemble algorithms—bagging, boosting, stacking, and majority voting—were employed in the experiments. A maximum accuracy increase of 7% for weak classifiers was achieved through ensemble classification. Additionally, the performance was further enhanced by implementing feature selection, resulting in significant improvements in prediction accuracy.

Swain et al. (2018) discussed that an accurate diagnosis is crucial, as it can save lives if the disease is identified early. ML classification models can assist medical practitioners in making decisions about diagnosing heart diseases. This survey analysed the performance of various heart disease prediction techniques, including ABC-SVM, ANFIS, SVM-ANN, SVM-SSVM, Genetic Algorithm, Neural Network Ensemble, FNN, and Majority Vote-Based Ensemble Classifier. These techniques were employed to analyse the Cleveland Heart Disease dataset from the UCI Machine Learning Repository.

Shah et al. (2020) introduced the problem of heart disease and highlighted the importance of accurate prediction for early intervention. It reviews existing literature on heart disease prediction and ML techniques. The authors described the dataset used from the Cleveland database of the UCI repository of heart disease patients. This research paper presents various attributes related to heart disease and the model based on supervised learning algorithms such as Naive Bayes, Decision Tree, K-nearest Neighbor, and Random Forest algorithms. The results show that the highest accuracy score was achieved with the K-nearest Neighbor algorithm.

Kavitha et al. (2021) proposed a novel ML approach to predict heart disease. The authors used the Cleveland heart disease dataset, employing data mining techniques such as regression and classification. ML techniques,

including Random Forest and Decision Tree, were applied. The model was designed using this novel method. The authors also implemented a hybrid model combining Random Forest and Decision Tree. Experimental results show an accuracy level of 88.7% for the heart disease prediction model using the hybrid approach.

Maini et al. (2020) explained an exploratory study conducted in South India that focuses on using machine learning algorithms to predict heart disease in the Indian population. The study utilized anonymized medical records from a tertiary hospital and applied five different state-of-the-art ML algorithms. The best-performing algorithm achieved a diagnostic accuracy of 93.8%. These promising results suggest that ML-based prediction systems could serve as a screening tool for diagnosing heart diseases in primary healthcare centers in India, where such diseases often go undetected. The article highlights the critical situation of cardiovascular diseases in India, which account for over 30% of total deaths.

Yang and Garibaldi (2015) discussed a hybrid model for automatically identifying risk factors for heart disease in clinical texts. The model employs a combination of machine learning, rule-based, and dictionary-based approaches to manage the complexity and variations in different types of risk factor evidence, including token-level clinical entities, sentence-level clinical facts, and sentence-level clinical measurements. The system achieved an overall micro-averaged F-measure of 0.915 on the i2b2 challenge test data, which was competitive with the best-performing system. The authors analysed the characteristics of the clinical evidence and categorized it into three main types. They discovered that the machine learning-based approaches work effectively for token-level clinical entity recognition, while rule-based and dictionary-based methods serve as beneficial supplements for sentence-level clinical facts and measurements.

Ali et al. (2019) introduce an expert system that combines two Support Vector Machine (SVM) models: one linear and L1 regularized, and the other L2 regularized, which are utilized as a predictive model. The hybrid grid search algorithm (HGSA) is proposed to optimize both models simultaneously. The effectiveness of the proposed method is evaluated using six metrics: accuracy, sensitivity, specificity, MCC, ROC charts, and AUC. Experimental results indicate that the proposed method enhances the performance of a conventional SVM model by 3.3%, surpassing ten previous methods and other state-of-the-art ML ensemble models.

Diwakar et al. (2020) explored the fusion of multiple modalities of medical data, such as combining imaging data with clinical data. The authors explained that this fusion can provide a more comprehensive understanding of a patient's condition, leading to more accurate predictions. ML models are being tailored to account for patient-specific characteristics, resulting in more personalized risk assessments and treatment plans. The authors demonstrate the integration of ML and image fusion techniques, which have significant potential to enhance the accuracy, efficiency, and personalized nature of heart disease prediction and diagnosis.

Keya et al. (2021) explore the use of ML algorithms to predict the likelihood of heart attacks. The paper discusses the significance of cardiovascular diseases, particularly heart attacks, and identifies risk factors such as age, smoking, high blood pressure, cholesterol, obesity, diabetes, and family history. It provides statistics on risk factors for heart disease, emphasizing the need for timely prediction and preventative measures. It compares five different algorithms—Logistic Regression, Random Forest, bagging, MLP, and Decision Tree in terms of accuracy and area under the curve (AUC). The study aims to identify the best algorithm for assessing the likelihood of a heart attack by utilizing correlation matrices, visualizing features, and analyzing AUC. The paper finds that Logistic Regression is the most effective model, achieving an accuracy of about 80% and the best AUC of approximately 87%.

Ansari et al. (2021) proposed a modified algorithm using logistic regression with principal component analysis to predict heart disease more accurately. In this paper, the UCI machine learning repository dataset is widely used for heart disease prediction systems. This model is evaluated on the UCI machine learning repository datasets to determine whether a person has heart disease. They proposed a model using logistic regression combined with principal component analysis. This led to the logistic regression model with all the variables, and the logistic regression model with PCA performed the best with an accuracy of 86%.

Pathan et al. (2022) focused on analysing the impact of feature selection techniques on the accuracy of heart disease prediction. They applied a filter-based feature selection technique, the ANOVA-F test, to identify the most relevant features from the datasets. The classification experiments were conducted using various ML models on both the complete and reduced feature sets. The study concludes that feature selection techniques can effectively identify critical risk factors for heart disease and improve overall prediction accuracy.

Louridi et al. (2021) discussed developing an effective intelligent medical system based on ML techniques to aid in identifying a patient’s heart condition and guide a doctor in accurately diagnosing whether a patient

has cardiovascular disease. They implemented multiple data processing techniques and addressed the issues of missing and imbalanced data in the publicly available datasets. The datasets used by the authors were the UCI Heart Disease dataset and the Framingham dataset. Additionally, they utilized XGboost, Adaboost, Gradient Boosting, Extra Trees, Light Gradient Boosting, SGDC, Nu SVM, and the stacking algorithm in the classification step, obtaining a score accuracy of 95.83% with the stacking algorithm.

Long et al. (2015) described a heart disease prediction system that uses a combination of rough sets based on attribute reduction and an interval type-2 fuzzy logic system (IT2FLS). The IT2FLS utilizes a hybrid learning process comprising a fuzzy c-means clustering algorithm and parameter tuning through chaos firefly and genetic hybrid algorithms to manage uncertainties in the datasets. Experiments on heart disease and SPECTF datasets show that the proposed system outperforms other machine learning methods like Naive Bayes, SVM, and ANN in terms of accuracy, convergence speed, and processing time.

Rani et al. (2021) proposed a system that describes several classification algorithms, including Naive Bayes, Support Vector Machine, Logistic Regression, Random Forest, and Adaboost. The Cleveland dataset from UCI was used to conduct experiments. Heart disease was diagnosed using various medical parameters available in the dataset. The authors likely collected relevant medical data such as patient demographics, medical history, diagnostic test results, and risk factors related to heart disease. The algorithms are trained and optimized with the collected data to predict the likelihood of heart disease in patients. The paper presents the experiment results, showcasing the predictive accuracy and effectiveness of the decision support system compared to traditional methods or individual ML algorithms.

Austin et al. (2013) compared the performance of modern classification and prediction methods from the data mining and ML literature with conventional classification and regression trees for classifying patients with heart failure (HF) into two subtypes: HF with preserved ejection fraction (HFPEF) and HF with reduced ejection fraction (HFREF). The authors found that modern, flexible tree-based methods such as bagged trees, Random Forests, and boosted trees provided a substantial improvement in predicting and classifying HF subtypes compared to conventional classification and regression trees.

The study by Chitra and Seenivasagam (2014) presents a method for early prediction of heart disease using a Neural Network optimized with particle swarm optimization (PSO). They use patient medical records and evaluate their PSO-optimized neural network (PSONN) with data from the Cleveland database and real-time clinical data. The optimized network parameters include the number of hidden neurons, the momentum factor, and the learning rate. The study concludes that the PSONN is an efficient and accurate tool for heart disease prediction, which could be a valuable aid for physicians.

Chauhan et al. (2018) discussed data mining and classification techniques for predicting heart disease. The study focuses on identifying an appropriate method that can aid in future decision-making. It analyzes the heart disease dataset using Decision Tree, Random Tree, and Random Forest classifiers. The goal is to establish a classifier that offers the highest accuracy for diagnosing the disease.

Mohan et al. (2019) proposed a novel method aimed at identifying significant features using machine learning techniques, thereby improving the accuracy of heart disease prediction. The model utilizes a hybrid random forest with a linear model (HRFLM) approach, combining features from Random Forest (RF) and Linear Method (LM). It achieves an accuracy level of 88.7%, demonstrating its potential for long-term savings of human life and the early detection of heart condition abnormalities. The study suggested that further research should focus on real-world datasets and develop new feature selection methods to enhance heart disease prediction performance.

Random forest swarm optimization-based approaches proposed by Asadi et al. (2021) were developed for diagnosing heart diseases. These approaches aim to select the optimal features that can enhance the accuracy of heart disease prediction. One such approach is the GAPSO-RF model, which combines a hybrid genetic algorithm (GA) and particle swarm optimization (PSO) with random forest (RF). Another approach is the PA-RF model, which utilizes a hybrid machine learning model based on random forest optimized by PSO and ant colony optimization (ACO). Additionally, a novel function for identifying optimal weights in the PSO algorithm has been introduced to improve feature selection and accuracy in heart disease diagnosis. These optimization-based approaches have demonstrated efficiency and robustness in classifying heart disease.

Maji and Arora (2019) discussed the use of data mining and machine learning techniques for predicting and diagnosing heart disease. The document describes a proposed hybrid model that combines C4.5 algorithms with ANN techniques to enhance the accuracy of heart disease prediction. The methodology involves preprocessing the dataset, applying the proposed hybrid technique, and comparing the performance

of the individual algorithms and the hybrid approach in terms of accuracy, sensitivity, and specificity. The document also addresses the applications of data mining in the healthcare industry, such as managing hospital resources, building relationships with customers, improving treatment techniques, and providing patient support and care.

Shilaskar and Ghatol (2013) discussed the use of feature selection techniques for medical diagnosis, particularly for cardiovascular diseases. It investigates the application of the Support Vector Machine for classification and compares various feature selection algorithms, including forward feature selection, backward elimination, and forward feature selection. The feature selection algorithms are evaluated based on accuracy, area under the curve (AUC), and the number of features selected. The results indicate that the proposed hybrid forward feature selection algorithm outperforms the other techniques, reducing the number of features while enhancing the accuracy of diagnosis.

Shah et al. (2020b) proposed a methodology based on a Support Vector Machine for heart disease diagnosis. They employed feature selection, wrapping selection, and feature extraction methods to enhance the performance of the Support Vector Machine model. The authors utilized feature subset selection techniques to reduce the dimensionality of the input space and improve the model's efficiency. The results indicate that the model, enhanced by feature subset selection, wrapping selection, and feature extraction methods, achieves better performance in terms of accuracy and other evaluation metrics compared to baseline models that lack these enhancements.

Singh et al. (2017) used supervised learning algorithms. The authors discussed ML algorithm techniques such as Linear Regression, Logistic Regression, Support Vector Machines, and Random Forest. The Cleveland heart disease dataset was utilized to apply these algorithms. The Cleveland Heart Disease dataset contains 74 attributes with 303 instances for each attribute, but only 14 of those attributes were selected. This paper employed optimal and effective 3-fold, 5-fold, and 10-fold cross-validation techniques to compare accuracy. The highest accuracy achieved is 85.81% using 10-fold cross-validation in the Random Forest Algorithm.

Khan et al. (2020) showed that 90% of cardiovascular diseases can be treated if predicted in advance. This study investigates the relationship between factors such as age, blood pressure, and gender concerning heart disease. Various classification models, including KNN, Decision Trees, Logistic Regression, Gaussian Naive Bayes, SVM, and Random Forests, are utilized to model the factors influencing heart disease. This experimental study aims to identify the best algorithm for categorizing factors related to heart disease diagnosis. The accuracy rate for the different algorithms was high, with Random Forest algorithms proving to be the most efficient.

An overview of ML classification techniques for heart disease prediction

This section presents a table of all the research papers described above. The comparison is based on accuracy and is displayed in Table 1. The table includes six elements, which are as follows:

i) Author: This indicates the author(s) of the paper and the reference number.

ii) Year: This shows the author(s) who published the paper of the year.

iii) Classification Techniques: This refers to the classification algorithm used in the research, including whether it involves a single algorithm, a comparison, or a hybrid model.

iv) Tool: This column indicates the framework or programming language used to construct the model. The researcher used it to preprocess the input dataset, create the predictive model, and conduct testing.

v) Dataset: This illustrates the dataset used as input for the classification algorithm.

vi) Accuracy: This reflects the accuracy of the proposed model's results.

Table 1. Comparison of classification techniques for heart disease prediction.

Author	Year	Classification Technique	Tool	Dataset	Accuracy
Gavhane et al.	2018	MLP	Python	Cleveland (UCI)	MLP (Precision 0.92)
Ambrish et al.	2022	LR	Python	UCI Dataset	87.10%
Dwivedi	2016	NB, KNN, LR, SVM, ANN, Classification Tree	Not Mentioned	UCI Statlog	85%
Latha and Jeeva	2019	Ensemble technique	WEKA	Cleveland (UCI)	Bagging & Boosting accuracy increases
Swain et al.	2018	ANN, SVM, RF	Not Mentioned	Cleveland (UCI)	ANN 97.5%
Shah et al.	2020a	NB, DT, KNN, RF	WEKA	Cleveland (UCI)	KNN 90.78%
Kavitha et al.	2021	DT, RF, and hybrid model using DT &	Python	Cleveland (UCI)	Hybrid Model 88.7%

Maini et al.	2020	RF KNN, NB, Logistic Regression, AdaBoost, and RF	Python	Hospital data from South India	RF 93.8%
Yang and Garibaldi	2015	ML with NLP Techniques	MALLET	University of Nottingham, UK	F-measure 0.915
Ali et al.	2016	Hybrid Grid Search Algorithm	Python	Cleveland (UCI)	HGSA improves the strength
Diwakar et al.	2020	NB, KNN, SVM, ANN	Not Mentioned	UCI Dataset	ANN has more accuracy than all other techniques,
Keya et al.	2021	Logistic Regression, Random Forest, bagging, MLP, and DT	Not Mentioned	Cleveland (UCI)	LR 80%
Ansari et al.	2021	LR, SVM, LR with PCA	Not Mentioned	Cleveland (UCI)	LR with PCA 86%
Pathan et al.	2022	LR, DT, RF, NB, MLP	Python	CVD & Framingham Dataset	MLP 0.75 for CVD Dataset
Louridi et al.	2021	XGboost, Adaboost, gradient boosting	Not Mentioned	UCI, Framingham Dataset	95.83%
Long et al.	2015	NB, SVM, ANN, Rough set with fuzzy logic	MATLAB, WEKA	UCI, SPECTF Dataset	Highest accuracy Rough set with fuzzy logic
Rani et al.	2021	NB, SVM, LR, RF, Adaboost	Python	Cleveland (UCI)	RF 86.6%
Austin et al.	2013	Classification & Regression	R programming	Clinical data in Ontario, Canada	Boosted trees had the best performance
Chitra and Seenivasagam	2014	SVM, MLFFNN, PSONN	Not Mentioned	UCI	PSONN 90.8%
Chauhan et al.	2018	DT, RF, Random Tree	Rapid miner	University of Lyon dataset	RF 75.60%
Mohan et al.	2019	LM and RF Hybrid HRFLM	R Studio Rattle	Cleveland (UCI)	HRFLM 88.7 %
Asadi et al.	2021	KNN, SVM, NB, RF, C4.5, LR and QDA.	MATLAB	Statlog, Cleveland, SPECT, SPECTF, VA Long Beach (UCI)	RF has more accuracy than other techniques.
Maji and Arora	2019	ANN, C4.5, Hybrid DT	WEKA	UCI	Hybrid DT 78.14%
Shilaskar and Ghatol	2013	SVM with Forward feature inclusion, Back-elimination, and Forward feature selection	Not Mentioned	UCI	Highest accuracy with Forward feature selection
Shah et al.	2020b	SVM	Not Mentioned	UCI- Cleveland, Hungarian, Switzerland	SVM 91.30%
Singh et al.	2017	Linear Regression, Logistic Regression, SVM, and RF	Python	Cleveland (UCI)	RF 85.81%
Khan et al.	2020	KNN, DT, Logistic Regression, Gaussian Naïve Bias, SVM, RF	Python	Cleveland (UCI)	RF 89%

Discussion

We categorized and surveyed numerous representative works on the application of ML classification algorithms. The tools utilized, the datasets used, the number of attributes and records, the preprocessing methods, and the classifiers included in the models all affect the accuracy of the proposed models.

The datasets used in the previous study include Cleveland, Hungarian, Switzerland, Statlog, SPECTF, Framingham, and hospital data. The software tools utilized are Python, MATLAB, WEKA, R Studio, Rattle, Rapid Miner, and R programming.

ML can be a powerful tool for predicting heart disease, potentially aiding healthcare professionals in early diagnosis and intervention. The algorithms utilized by researchers include ANN, Adaboost, DT, Linear Regression, Logistic Regression, Naive Bayes, SVM, RF, MLP, hybrid models, and more.

A dataset with sufficient samples and accurate data must be used to create a precise heart disease prediction model. The most crucial step in preparing the dataset for the ML algorithm is to preprocess it appropriately to ensure acceptable results.

Conclusion

One of the most critical fields in the medical sector is heart disease prediction, which utilizes available patient data to detect the presence or absence of heart disease. Various ML algorithms exist for heart disease prediction, and preprocessing methods are employed to enhance their performance. The most crucial step is

feature selection, which improves the accuracy of the algorithms. ML holds significant potential for forecasting heart disease. Studies have achieved high accuracy rates using a variety of algorithms to assess patient data. This tool could be beneficial for doctors, leading to earlier detection, improved prevention efforts, and better patient outcomes. ML models are only as effective as the data on which they are trained. More data and higher data quality can result in more accurate models. In the future, we plan to improve accuracy by combining ML classifier. The implemented algorithms can be enhanced by integrating multiple techniques and creating a hybrid system to boost performance and efficiency.

References

Ali, L., Niamat, A., Khan, J. A., Golilarz, N. A., Xingzhong, X., Noor, A., Nour, R., & Bukhari, S. A. C. (2019). An optimized stacked support vector machine-based expert system for the effective prediction of heart failure. *IEEE Access*, *7*, 54007-54014. https://doi.org/10.1109/ACCESS.2019.2909969

Ambrish, G., Ganesh, B., Ganesh, A., Srinivas, C., Dhanraj, & Mensink, K. (2022). Logistic regression technique for the prediction of cardiovascular disease. *Global Transitions Proceedings*, *3*(1), 127-130. https://doi.org/10.1016/j.gltp.2022.04.008

Ansari, M. F., Kaur, B. A., & Kaur, H. (2021). A prediction of heart disease using machine learning algorithms. In J. IZ. Chen, J. M. R. S. Tavares, S. Shakya, & A. M. Iliyasu (Eds), *Image processing and capsule networks* (pp.497-504). Springer. https://doi.org/10.1007/978-3-030-51859-2_45

Asadi, S., Roshan, S., & Kattan, M. W. (2021). Random forest swarm optimization-based for heart disease diagnosis. *Journal of Biomedical Informatics*, *115*, 1-13. https://doi.org/10.1016/j.jbi.2021.103690

Austin, P. C. Tu, J. V., Ho, J. E., Levy, D., & Lee, D. S. (2013). Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *Journal of Clinical Epidemiology*, *66*, 398-407. https://doi.org/10.1016/j.jclinepi.2012.11.008

Chauhan, R., Jangade, R., & Rekapally, R. (2018). Classification model for prediction of heart disease. In M. Pant, K. Ray, T. Sharma, S. Rawat, & A. Bandyopadhyay (Eds.), *Soft computing: Theories and applications. Advances in intelligent systems and computing* (v. 584) (pp.707-714). Springer. https://doi.org/10.1007/978-981-10-5699-4_67

Chitra, R., & Seenivasagam, V. (2014). Risk prediction of heart disease based on swarm optimized neural network. In S. Patnaik, & X. Li (Eds), *Proceedings of International Conference on Computer Science and Information Technology* (v. 255) (pp. 707-714). Springer. https://doi.org/10.1007/978-81-322-1759-6_81

Diwakar, M., Tripathi, A., Joshi, K., Memoria, M., Singh, P., & Kumar, N. (2020). Latest trends on heart disease prediction using machine learning and image fusion. *Materialstoday Proceedings*, *37*(2), 3213-3218. https://doi.org/10.1016/j.matpr.2020.09.078

Dwivedi, A. K. (2016). Performance evaluation of different machine learning techniques for the prediction of heart disease. *Neural Computing & Applications*, *29*, 685-693. https://doi.org/10.1007/s00521-016-2604-1

Gavhane, A., Kokkula, G., Pandya, I., & Devadkar, K. (2018). Prediction of heart disease using machine learning. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 1275-1278). IEEE. https://doi.org/10.1109/ICECA.2018.8474922

Kavitha, M., Gnaneswar, G., Dinesh, R., Rohith Sai, Y., & Sai Suraj, R. (2021). Heart disease prediction using hybrid machine learning model. *In International Conference on Inventive Computation Technologies* (pp. 1329-1333). IEEE. https://doi.org/10.1109/ICICT50816.2021.9358597

Keya, M. S., Shamsojjaman, M. Hossain, F., Akter, F., Islam, F., & Emon, M. U. (2021). Measuring the heart attack possibility using different types of machine learning algorithms. In *2021 International Conference on Artificial Intelligence and Smart Systems* (pp. 74-78). IEEE. https://doi.org/10.1109/ICAIS50930.2021.9395846

Khan, Z., Mishra, D. K., Sharma, V., & Sharma, A. (2020). Empirical study of various classification techniques for heart disease prediction. In *International conference on computing, communication and automation* (pp. 57-62). IEEE. https://doi.org/10.1109/ICCCA49541.2020.9250852

Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of the prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, 1-9, https://doi.org/10.1016/j.imu.2019.100203

Long, N. C., Meesad, P., & Unger, H. (2015). A highly accurate firefly-based algorithm for heart disease prediction. *Expert Systems with Applications*, 42(21), 8221-8231. https://doi.org/10.1016/j.eswa.2015.06.024

Louridi, N., Douzi, S., & El Ouahidi, B. (2021). Machine learning-based identification of patients with a cardiovascular defect. *Journal of Big Data*, 8(133), 1-15. https://doi.org/10.1186/s40537-021-00524-9

Maini, E., Venkateswarlu, B., Maini, B., & Marwaha, D. (2020). Machine learning based heart disease prediction system for Indian population: An exploratory study done in South India. *Medical Journal Armed Forces India*, *77*(3), 302-311. https://doi.org/10.1016/j.mjafi.2020.10.013

Maji, S., & Arora, S. (2019). Decision tree algorithms for prediction of heart disease. In S. Fong, S. Akashe, & P. Mahalle (Eds.), *Information and communication technology for competitive strategies. Lecture notes in networks and systems* (v. 40) (pp. 447-454). Springer. https://doi.org/10.1007/978-981-13-0586-3_45

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, *7*, 81542-81554. https://doi.org/10.1109/ACCESS.2019.2923707

Pathan, M. S., Nag, A., Pathan, M. M. & Dev, S. (2022). Analyzing the impact of feature selection on the accuracy of heart disease prediction. *Healthcare Analytics*, *2*, 1-9. https://doi.org/10.1016/j.health.2022.100060

Rani, P., Kumar, R., Ahmed, N. M. O. S., & Jain, A. (2021). A decision support system for heart disease prediction based upon machine learning. *Journal of Reliable Intelligent Environments*, *7*, 263–275. https://doi.org/10.1007/s40860-021-00133-6

Shah, D., Patel, S., & Bharti, S. K. (2020a). Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*(345). https://doi.org/10.1007/s42979-020-00365-y

Shah, S. M. S., Shah, F. A., Hussain, S. A., & Batool, S. (2020b). Support vector machines-based heart disease diagnosis using feature subset, wrapping selection and extraction methods. *Computers and Electrical Engineering*, *84*, 106628. https://doi.org/10.1016/j.compeleceng.2020.106628

Shilaskar, S., & Ghatol, A. (2013). Feature selection for medical diagnosis: Evaluation for cardiovascular diseases. *Expert Systems with Applications*, *40*(10), 4146-4153. https://doi.org/10.1016/j.eswa.2013.01.032

Singh, Y. K., Sinha, N., & Singh, S. K. (2017). Heart disease prediction system using random forest. In M. Singh, P. Gupta, V. Tyagi, A. Sharma, T. Ören, & W. Grosky (Eds.), *Advances in computing and data sciences*. *Communications in computer and information science* (v. 721) (pp.613-623). Springer. https://doi.org/10.1007/978-981-10-5427-3_63

Swain, D., Pani, S. K., & Swain, D. (2018). A metaphoric investigation on prediction of heart disease using machine learning. In *International Conference on Advanced Computation and Telecommunication* (pp. 1-6). IEEE. https://doi.org/10.1109/ICACAT.2018.8933603

World Health Organization [WHO]. (2020). *Cardiovascular diseases*. https://www.who.int/healthtopics/cardiovascular-diseases.

Yang, H., & Garibaldi, J. M. (2015). A hybrid model for automatic identification of risk factors for heart disease. *Journal of Biomedical Informatics*, *58*(Suppl.), 171-182. https://doi.org/10.1016/j.jbi.2015.09.006