*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 12 Issue: 2*
***Article Received:*** *25 November2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

# Prediction of Heart Disease Using Machine Learning Techniques

**Mohammed Jasim A.Alkhafaji**[1]
[1]Computer Technology Engineering
Al_Taff university college
Karbala, Iraq

**Samah A Nasir**[1,2]
[1]Computer Technology Engineering, Al_Taff university college
[2]Ministry of Education, Karbala Directorate
Kerbala, Iraq

**Zinah alhussein**[1]
[1]Computer Technology Engineering
Al_Taff university college
Karbala, Iraq

**Abstract—** A potential strategy in the healthcare industry is the prediction of cardiac disease using machine learning algorithms. Worldwide, heart disease continues to be one of the major causes of death, and successful treatment and prevention depend greatly on early identification. Large volumes of patient data may be analyzed using machine learning algorithms to find patterns and risk factors that might lead to the onset of heart disease. These algorithms use supervised learning, unsupervised learning, and ensemble approaches to assess a variety of data sources, including clinical test results, patient demographics, and medical records. Machine-learning algorithms may be trained on historical data from a variety of patients to discover complicated associations and generate precise predictions about a person's risk of acquiring heart disease. Our objective is to create a machine-learning technique that reliably predicts heart disease and is computationally effective. Feature selection is a crucial step in the creation of prediction models as it permits the identification of the most significant risk factors for heart disease. Machine learning methods including logistic regression, support vector machines, decision trees, random forests, and neural networks are often used to predict cardiac disease. By examining extensive patient data, machine learning algorithms show considerable potential in the prediction of cardiac disease. In the battle against heart disease, their capacity to spot patterns and risk factors may result in early identification, individualized therapies, and better patient care.

**Keywords-** data mining, Prediction, heart disease, machine learning, technique.

## I. INTRODUCTION

The World Health Organization (WHO) has a crucial role to play in tackling the problem of heart disease prediction, which is a major area of concern for global health. The WHO is a specialized department of the UN that deals with global public health. It works to combat various diseases, including cardiovascular diseases like heart disease, through research, policy development, and global coordination efforts. The WHO recognizes heart disease as a major global health challenge and emphasizes the importance of prevention, early detection, and treatment. It collaborates with member countries to develop strategies and guidelines for heart disease prevention and control. The organization also promotes awareness campaigns to educate the public about risk factors and healthy lifestyle choices that can reduce the burden of heart disease [1].

heart disease mortality, remains a leading cause of death worldwide. According to the WHO's latest global health estimates from 2020, cardiovascular diseases accounted for approximately 17.9 million deaths or 32% of all deaths globally. Among cardiovascular diseases, ischemic heart disease (caused by narrowed coronary arteries) is the leading cause, responsible for 8.9 million deaths, followed by stroke with 6.3 million deaths. The World Health Organization (WHO) recognizes the importance of various medical data in predicting heart disease. These data points provide valuable insights into an individual's risk factors and help in assessing the likelihood of developing cardiovascular conditions [2]. Information about previous cardiovascular events, such as heart attacks, strokes, or other heart-related conditions, helps assess an individual's predisposition to heart disease. The presence of known risk factors, including hypertension (high blood pressure), dyslipidemia (abnormal cholesterol levels), diabetes, smoking, obesity, and family history of heart disease, significantly contribute to the prediction of heart disease [3]. Symptoms like chest pain (angina), shortness of breath, palpitations, and fatigue, along with clinical assessments such as blood pressure measurements, heart rate, and electrocardiogram (ECG) findings, provide important diagnostic information. Measurements of blood markers, such

as lipid profiles (HDL cholesterol, total cholesterol, LDL cholesterol, and triglycerides), blood glucose levels, and other biomarkers like C-reactive protein (CRP), can assist in assessing cardiovascular risk [4]. Diagnostic tests like echocardiography, stress tests (exercise or pharmacological), coronary angiography, and cardiac computed tomography (CT) scans can provide visual evidence of structural abnormalities, blockages in blood vessels, and other cardiac conditions. These medical data, combined with demographic information and lifestyle factors, form a comprehensive profile that machine-learning algorithms can analyze to predict heart disease risk [5].

**2- Literature survey**

The Cleveland heart disease database, which is publicly accessible online at a UCI data-mining repository, has been the subject of several research to assess the classification accuracy of different machine learning techniques. On this dataset, the authors of [6] were able to achieve a prediction accuracy of 77% by utilizing the logistic regression technique. In this study, authors [7] enhanced their work and noticed better prediction accuracy by contrasting several global evolutionary computing techniques. a calm Work on using machine learning algorithms to identify cardiovascular heart disease has had a major effect on this research. In this article, a summary of the literature is presented. Several algorithms, including Logistic Regression, KNN, Random Forest Classifier, and others, have been used to create a trustworthy prediction of cardiovascular disease. The results demonstrate that each algorithm has a distinct capacity to achieve the predetermined aims [8]. Authors Bayu Adhi Tama, et al. [9] suggested research on the detection of diabetes illness using machine learning methods in their paper. This condition was considered to be a crucial element of ML. Over 285 million people worldwide have diabetes, according to a study by the International Diabetes Federation (IDF). Although it is difficult to identify type 2 diabetes early, the author's research, which employed data mining since it yields the best results, helped to reveal information from publicly accessible data. In their investigation, researchers mined historical records for relevant information about specific patients using SVMs. Patients were able to get the proper therapy and lower their risk of complications because of an early identification of type 2 diabetes. ANN has been introduced [10]. To provide the medical field forecasts with the maximum accuracy possible. The backpropagation multilayer perception (MLP) of ANN is used to forecast heart disease. The outcomes are compared to other models that have been used in the same field, and it is shown that they are superior [11]. The data of heart disease patients collected from the UCI laboratory is searched for patterns using NN, DT, SVM, and Naive Bayes.

A variety of applications examined by Yu-Xuan Wang et al. [12]. have shown the usefulness of ML approaches in a wide range of domains. They proposed a fresh approach to creating a practical framework. Several machine-learning methods were included in the plan. When the data miner yielded the desired result, all the information obtained from the structure was reviewed. The various tests showed that the proposed technique delivered outstanding results. Zhiqiang Ge et al. provided an earlier article on applications for analytics and data mining in 2017. Forsome reasons, these procedures were used in the corporate sector. Here, they have looked at 8 unsupervised learning techniques and 10 supervised learning algorithms [13]. In their research, they showed how semi-supervised type learning algorithms may be used. According to industry approaches, supervised and unsupervised machine learning techniques were deployed in between 90% and 95% of applications. As a consequence, it was recommended that the creation of various novel applications in fields like medicine requires the use of machine learning methods.

**3- Dimensionality Reduction**

To include just the most important information, dimension reduction entails choosing a mathematical representation that allows one to relate the majority, but not all, of the variation within the provided data. There may be many features or dimensions in the data being examined for a job or an issue, but not all of these attributes will have an equal impact on the outcome. The computational complexity may be impacted by a high number of qualities or features, and it may even result in overfitting, which yields subpar results. Dimensionality Reduction is thus a crucial phase that must be taken into account while creating any model. Two techniques are often used to reduce dimensions: feature Choice and extraction of features.

**A. Extraction of Features**

In this, a new set of features is produced using the first feature set. Before being retrieved, the attributes must be changed. It is often hard to undo the change since some, maybe a lot, of crucial information is lost in the process. For feature extraction, Principal Component Analysis (PCA) is employed in [14] and [15]. Principal component analysis is one of the often-used linear transformation algorithms. In the feature space, it searches for routes that maximize variance as well as pathways that are orthogonal to one another. The best reconstruction is achieved using a global approach.

**B. Feature Choice**

This selects a piece of the first feature set. The CFS (Correlation-based Feature Selection) Subset Evaluation strategy is combined with the Best First Search method to reduce dimensionality [16]. To choose the most important attributes, [17] chi-square statistics test is employed.

**4- Objectives**

Data mining strategies must include data pre-processing, which comprises understandably placing raw data. Real-world medical information is often incomplete and erroneous for certain behaviors or trends. Additionally, it is often insufficient and unreliable. A tried-and-true method to address these issues is preprocessing data. Data pre-processing is the procedure that gets raw data ready for further processing.

The major goal of this project is to provide an intuitive platform where patients may submit their medical information, and an algorithm will be used to determine the kind of heart disease **based** on the characteristics gathered. It also saves time and makes it easier for patients and doctors to predict a

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 12 Issue: 2*
***Article Received:*** *25 November2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

patient's tendency for any kind of heart illness, which is often difficult to accomplish without a doctor's help. As this algorithm completes the process, a well-trained model is less likely to err in predicting the heart illness and its type.
To diagnose cardiac disease, other medical information, such as age, cholesterol, etc., must be provided. Since there is no human participation, the odds of mistakes are quite minimal. The algorithm will then display the findings based on the characteristics gathered. Additionally, it saves a lot of time for patients or medical professionals, enabling them to complete treatments or other operations more rapidly. This is true if they get their results earlier. The precaution/prevention phase of heart therapy may progress more quickly if doctors and patients are allowed to use this crucial time on alternative therapies and preventive measures that will decrease the impact of heart disease.

## 5. Methodology

Developing a heart disease prediction model typically involves the following main steps.

*A. Data Collection: Gather a heart disease dataset that includes relevant features such as patient demographics, medical history, lifestyle factors, symptoms, and results from diagnostic tests (e.g., electrocardiogram and blood tests). Ensure the dataset includes labeled examples indicating the presence or absence of heart disease. Sources for heart disease datasets include medical research repositories, public health databases, and hospitals. In our study, we utilized the Cleveland Heart Disease Dataset, which is available online at the UCI Repository [18].*

| S. No. | Attribute | Description |
|---|---|---|
| 1 | Id | Id number |
| 2 | Sex | male, female |
| 3 | Age | In year |
| 4 | chest pain | Angina, abnang, notang, asympt. |
| 5 | Smoke | yes=1, no=0 |
| 6 | FBS | 0 if <120 mg/dl, 1 if >120 mg/dl. |
| 7 | Induced angina | no=0, yes=1. |
| 8 | ECG | Resting electrocardiographic Results |
| 9 | The slope | slope of the ST segment |
| 10 | THALACH | Maximum Heart Rate observed |

Table 1. Features

B. Data preprocessing: To manage outliers, inconsistencies, and missing values, outliers, clean and preprocess the dataset. To maintain compatibility across features, do data transformations such as normalization or feature scaling. To evaluate the model, divide the dataset into training and testing sets. Consideration should be given to maintaining data integrity and privacy [19].
C. Feature Selection/Engineering: Identify relevant features for the prediction of heart disease by analyzing the dataset. Highlight choice procedures like relationship examination, factual tests (e.g., chi-square), or element significance from calculations (e.g., arbitrary timberlands) can assist with distinguishing educational elements. Also, include designing might include making new elements or changing existing ones to improve predictive power.
D. Model Selection: Pick a reasonable AI algorithm for coronary illness expectation. Generally utilized calculations incorporate strategic regression, irregular backwoods, decision trees, neural networks, and support vector machines (SVM). Consider factors such as interpretability, computational efficiency, and performance on similar datasets. Domain knowledge and literature review can guide the selection process.
E. Model Training: Using the training dataset, train the selected model. As cardiac disease is present or absent, the model learns patterns and correlations between the characteristics. To reduce prediction errors, the model's parameters are optimized throughout the training phase. Cross-validation is one method that may be used to evaluate model performance and avoid overfitting.
F. Model Evaluation: Assess how well the tested dataset performs the training model. Analyze measures including area under the ROC curve (AUC-ROC), recall, accuracy, precision, and F1 score. An understanding of the model's capacity to generalize to unknown data is ensured through proper assessment. Think about using medical standards or domain-specific benchmarks for performance evaluation [20] [21].
G. Model Optimization: Calibrate the model's hyperparameters to work on its exhibition. Strategies like matrix search, irregular inquiry, or Bayesian improvement can help recognize ideal hyperparameter arrangements. Regularization procedures like L1 or L2 regularization can likewise be utilized to stay away from overfitting.
H. Model Validation: Validate the model's performance using an independent dataset, preferably from a different source or period, time. This step confirms the model's generalizability and robustness.
I. Model Deployment: Integrate the heart disease prediction model into a real-world application or healthcare system. Ensure that the deployment process considers factors such as scalability, security, and compliance with relevant regulations (e.g., patient data privacy) [22].

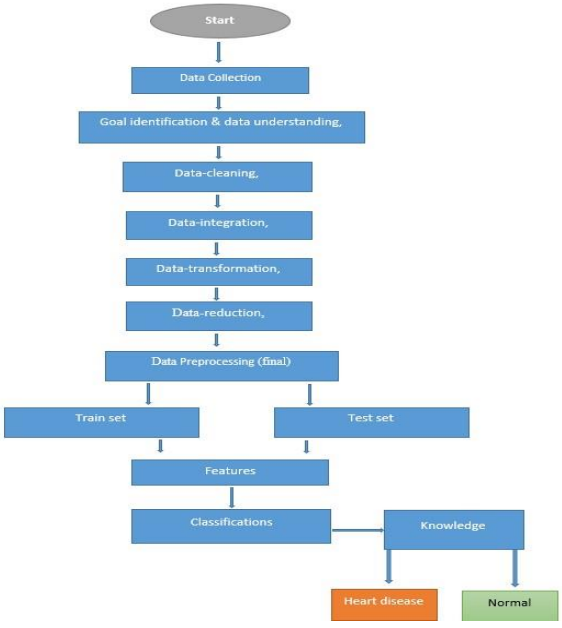


Figure 1. Proposed system

## 6. Result & discussion and future attention

Machine learning algorithms and techniques have demonstrated significant power and effectiveness in predicting heart disease. A broad variety of patient data, including demographics, medical history, symptoms, and results from diagnostic tests, may be automatically sorted through to find the most relevant aspects. This helps in identifying the key factors that contribute to heart disease and enables a better understanding and interpretation of the disease's risk factors. Machine learning algorithms are capable of capturing complex patterns and interactions between different features, including non-linear relationships that may not be easily discernible through traditional statistical methods. This enables the identification of subtle relationships between variables and provides insights into the risk factors and predictors of heart disease. Machine learning algorithms can achieve high accuracy in predicting heart disease. To learn from a lot of data and create precise predictions, they use sophisticated mathematical models and optimization approaches. Decision trees, random forests, support vector machines (SVM), logistic regression, and neural networks are a few examples of algorithms that have shown high performance in heart disease prediction tasks. Machine learning algorithms can stratify individuals into different risk categories based on their likelihood of developing heart disease. By analyzing multiple risk factors and patient characteristics, these algorithms can provide personalized risk assessments and identify individuals who may benefit from preventive interventions or early detection. Machine learning techniques can aid in the early detection and diagnosis of heart disease by leveraging patterns and markers in patient data. They can analyze various data sources such as electrocardiograms (ECG), echocardiograms, laboratory tests, and medical imaging to assist in accurate and timely diagnosis, potentially leading to improved treatment outcomes. Machine learning algorithms can effectively integrate and analyze data from various sources, such as electronic health records (EHRs), wearable devices, and genetic data. This integration allows for a comprehensive assessment of an individual's health status and facilitates more accurate predictions and risk assessments. Machine learning algorithms can continuously learn and adapt to new data and evolving patterns. As more data becomes available and medical knowledge advances, these algorithms can be updated to incorporate the latest information, improving their predictive performance over time.

| Model | Result |
|---|---|
| SVM | 98% |
| decision tree | 88% |
| Naïve Bayes | 82% |
| Random Forest | 88% |
| KNN | 97% |
| Logistic regression | 95% |

Table 2. Results of the most popular heart disease prediction algorithms and methods.

These results show that KNN, Random Forest Classifier, and Logistic Regression deliver better results to patients with Heart Disease, despite the fact that the bulk of research employs other algorithms, such as SVC, and decision Tree, to identify them. The highest accuracy generated by KNN and Logistic Regression is also more or almost comparable to the accuracy generated by preceding experiments. In conclusion, our accuracy has increased as a result of the extra medical features we used from the dataset we obtained. Additionally, our experiment demonstrates that Logistic Regression and KNN outperform Random Forest Classifiers in terms of predicting whether a patient will be diagnosed with heart disease. This shows that KNN and Logistic Regression are more effective in diagnosing heart disease.

It is worth noting that while machine-learning algorithms have shown promise in predicting heart disease, their clinical deployment should involve careful validation, interpretation, and consideration of ethical considerations. Collaboration between machine learning experts and healthcare professionals is crucial for the successful implementation and effective utilization of these algorithms.

## 7. Conclusion

Given the aforementioned study, machine learning algorithms show significant potential for forecasting cardiovascular disorders and illnesses that are connected to the heart. Every algorithm listed above has worked successfully in some circumstances and ineffectively in others. When paired with PCA, exchanging decision trees has shown exceptionally good performance; nevertheless, in other scenarios, choice trees have shown relatively poor performance, which may be related to overfitting. Both Logistic Regression and KNN gave excellent performances. The models that relied on the naive Bayes classifier had good performance and low computational overhead. SVM excelled in the vast majority of situations. Despite the fact that systems based on machine learning algorithms and approaches have shown to be highly effective in predicting heart-related disorders, more study is needed to determine the best way to deal with high-dimensional data and overfitting. The optimal selection of algorithms to use with a certain kind of data may also be thoroughly researched.

## REFERENCES

[1] W. H. Organization, *Prevention of cardiovascular disease: guidelines for assessment and management of total cardiovascular risk*. World Health Organization, 2007.

[2] W. H. Organization, “Risk reduction of cognitive decline and dementia: WHO guidelines,” 2019.

[3] M. J. A. Alkhafaji, A. F. Aljuboori, and A. A. Ibrahim, “Clean medical data and predict heart disease,” in *HORA 2020 - 2nd International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings*, 2020. doi: 10.1109/HORA49412.2020.9152870.

[4] G. A. Mensah, G. A. Roth, and V. Fuster, “The global burden of cardiovascular diseases and risk factors: 2020 and beyond,” *Journal of the American College of Cardiology*, vol. 74, no. 20. American College of

Cardiology Foundation Washington, DC, pp. 2529–2532, 2019.

[5] W. H. Organization, *Global status report on noncommunicable diseases 2010*. World Health Organization, 2011.

[6] R. Detrano *et al.*, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *Am. J. Cardiol.*, vol. 64, no. 5, pp. 304–310, 1989.

[7] B. Edmonds, "Using localised 'Gossip'to structure distributed learning," in *AISB*, Citeseer, 2005, pp. 127–134.

[8] A. Ganna *et al.*, "Multilocus genetic risk scores for coronary heart disease prediction," *Arterioscler. Thromb. Vasc. Biol.*, vol. 33, no. 9, pp. 2267–2272, 2013.

[9] B. A. Tama, A. Firdaus, and F. S. Rodiyatul, "Detection of type 2 diabetes mellitus disease with data mining approach using support vector machine," in *Proceeding of The 2010 International Conference on Informatics, Cybernetics, and Computer Applications. Bangalore, India*, 2010.

[10] L. Baccour, "Amended fused TOPSIS-VIKOR for classification (ATOVIC) applied to some UCI data sets," *Expert Syst. Appl.*, vol. 99, pp. 115–125, 2018.

[11] 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.

[12] Y.-X. Wang, Q. Sun, T.-Y. Chien, and P.-C. Huang, "Using data mining and machine learning techniques for system design space exploration and automatized optimization," in *2017 International Conference on Applied System Innovation (ICASI)*, IEEE, 2017, pp. 1079–1082.

[13] Z. Ge, Z. Song, S. X. Ding, and B. Huang, "Data mining and analytics in the process industry: The role of machine learning," *Ieee Access*, vol. 5, pp. 20590–20616, 2017.

[14] B. D. Kanchan and M. M. Kishor, "Study of machine learning algorithms for special disease prediction using principal of component analysis," in *2016 international conference on global trends in signal processing, information computing and communication (ICGTSPICC)*, IEEE, 2016, pp. 5–10.

[15] R. Kavitha and E. Kannan, "An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining," in *2016 international conference on emerging trends in engineering, technology and science (icetets)*, IEEE, 2016, pp. 1–5.

[16] S. Xu, Z. Zhang, D. Wang, J. Hu, X. Duan, and T. Zhu, "Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework," in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, IEEE, 2017, pp. 228–232.

[17] M. Singh, L. M. Martins, P. Joanis, and V. K. Mago, "Building a cardiovascular disease predictive model using structural equation model & fuzzy cognitive map," in *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, IEEE, 2016, pp. 1377–1382.

[18] "No Title", [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+disease

[19] A. Asuncion and D. Newman, "UCI machine learning repository." Irvine, CA, USA, 2007.

[20] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[21] M. Aljanabi, M. H. Qutqut, and M. Hijjawi, "Machine learning classification techniques for heart disease prediction: a review," *Int. J. Eng. Technol.*, vol. 7, no. 4, pp. 5373–5379, 2018.

[22] Y. Zhao, E. P. Wood, N. Mirin, S. H. Cook, and R. Chunara, "Social determinants in machine learning cardiovascular disease prediction models: a systematic review," *Am. J. Prev. Med.*, vol. 61, no. 4, pp. 596–605, 2021.