International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Review Article

Review on Heart Disease Prediction using Machine Learning

Nidhi Singh[1*] and Bhanu Pratap Singh[2]

Computer Science & Engineering, VITS Engineering College, Satna, M.P., India

Received 26 March 2025, Accepted 20 April 2025, Available online 23 April 2025, **Vol.15, No.2 (March/April 2025)**

Abstract

Cardiovascular disease remains the leading cause of death across the globe. Predicting heart disease (HDP) is a complex task that demands specialized knowledge and experience. Additionally, it faces numerous challenges related to clinical data analysis. Despite extensive research in this area, prediction accuracy—a key performance metric—still falls short of expectations. Accurate HDP can be life-saving, while incorrect predictions may lead to fatal consequences. To address these concerns, this review explores various Heart Disease Prediction techniques based on Deep Learning (DL), Machine Learning (ML), and optimization methods. Recently, numerous researchers have applied DL and ML algorithms to support healthcare professionals in diagnosing heart disease. The paper also examines different optimization algorithms and their performance. Ultimately, this review suggests that optimization-based HDP methods can play a crucial role in helping doctors predict heart disease early and recommend appropriate treatments.

Keywords: *Machine Learning, Supervised learning, Support vector machine, Random Forest, Neural Network.*

Introduction

The heart is a vital organ in the human body, responsible for pumping blood to every part of our system. If it fails to function properly, essential organs like the brain cease to work, leading to death within minutes. Unhealthy lifestyle choices, work-related stress, and poor dietary habits have significantly contributed to the rise in heart-related illnesses. Today, heart disease stands as one of the leading causes of death globally. According to the World Health Organization (WHO), cardiovascular diseases claim around 17.7 million lives annually, accounting for 31% of all global deaths. In India, too, heart disease has become the primary cause of mortality. The 2023 Global Burden of Disease Report, revealed that heart-related ailments, or cardiovascular diseases (CVDs), claimed an estimated 20.5 million lives in 2021. Beyond the human cost, these diseases also lead to increased healthcare expenditures and reduced productivity. WHO estimates indicate that India lost nearly $337 billion between 2010 and 2020 due to cardiovascular diseases. Therefore, the ability to accurately and effectively predict heart-related conditions is of critical importance.

Medical organizations worldwide gather vast amounts of data related to various health conditions. This data can be effectively analyzed using machine learning techniques to uncover valuable insights. However, due to the sheer volume and potential noisiness of the data, it often becomes difficult for humans to interpret it manually. Machine learning algorithms are particularly well-suited for handling such complex and large-scale datasets. As a result, these techniques have recently become highly effective tools for accurately predicting the presence or absence of heart-related diseases. The adoption of information technology in the healthcare sector is steadily growing, providing valuable support to doctors in their decision-making processes. It assists medical professionals in managing diseases, prescribing medications, and uncovering patterns and relationships within diagnostic data. Traditional methods for predicting cardiovascular risk often fall short, failing to identify many individuals who could benefit from preventive care, while also subjecting others to unnecessary treatments. Machine learning presents a promising solution by leveraging complex interactions among various risk factors to enhance prediction accuracy. This study evaluates whether machine learning can improve the prediction of cardiovascular risk.

2. Literature survey

Chala Beyene *et al*. [1] proposed a method for predicting and analyzing the occurrence of heart disease using data mining techniques. The primary goal of their study is to enable early and automated diagnosis of heart disease, delivering results in a short period. This approach is particularly valuable in

*Corresponding author's ORCID ID: 0000-000-0000-0000
DOI: https://doi.org/10.14741/ijcet/v.15.2.13

healthcare settings where medical professionals may lack extensive expertise. The methodology incorporates various medical attributes—such as blood sugar levels, heart rate, age, and sex—to determine whether an individual is at risk of heart disease. The dataset analysis was conducted using the WEKA software.

Senthilkumar Mohan *et al*. [2] implemented a hybrid machine learning approach for predicting heart disease using the Cleveland dataset. The process begins with data preprocessing, where tuples with missing values are removed. The authors excluded attributes like age and sex, considering them personal and not impactful on prediction accuracy. Instead, they focused on the remaining 11 attributes, which contain essential clinical information. They introduced a novel method called Hybrid Random Forest Linear Method (HRFLM), which combines Random Forest (RF) and Linear Method (LM). The HRFLM algorithm comprises four main steps. The first algorithm partitions the dataset using decision trees, executing them for each sample to segment the data into leaf nodes. The output is a partitioned dataset. In the second step, classification rules are applied to these partitions, resulting in labeled data. The third algorithm involves feature extraction using a Less Error Classifier, which identifies features by minimizing classification error rates. The final step applies a hybrid classifier that evaluates extracted features based on error rates to make predictions. The performance of HRFLM was compared with other classification techniques like decision trees and support vector machines. Since RF and LM individually showed strong performance, combining them into the HRFLM led to even better results. The authors conclude that further improvements in prediction accuracy could be achieved by integrating various machine learning algorithms.

Ali, Liaqat, *et al*. [3] proposed a system that includes two models based on linear Support Vector Machines (SVM). The first model employs L1 regularization to eliminate irrelevant features by reducing their coefficients to zero, while the second model uses L2 regularization for the actual prediction of heart disease. To fine-tune both models, the authors introduced a hybrid grid search algorithm that optimizes performance based on several evaluation metrics, including accuracy, sensitivity, specificity, the Matthews correlation coefficient, the ROC curve, and the area under the curve (AUC). The Cleveland dataset was used for experimentation, with a 70% training and 30% testing split via holdout validation. Two separate experiments were conducted, each exploring different values for the hyperparameters C1, C2, and k—where C1 is for the L1-regularized model, C2 for the L2-regularized model, and k represents the number of selected features. In the first experiment, the L1-linear SVM model was stacked with the L2-linear SVM model, achieving a maximum testing accuracy of 91.11% and a training accuracy of 84.05%. The second experiment combined the L1-linear SVM model with an L2-linear SVM model using an RBF kernel, yielding even better results—with a testing accuracy of 92.22% and training accuracy of 85.02%. Overall, their approach demonstrated a 3.3% improvement in accuracy over traditional SVM models. Singh, Yeshvendra K. *et al*. [4] explored several supervised machine learning algorithms—such as Random Forest, Support Vector Machine (SVM), Logistic Regression, Linear Regression, and Decision Tree—for heart disease prediction. They used the Cleveland dataset, which originally contained 303 tuples, of which six had missing values. These were removed during preprocessing, leaving 297 valid entries. The dataset was then split into 70% for training and 30% for testing. They applied multiple validation techniques, including 3-fold, 5-fold, and 10-fold cross-validation. The first algorithm used was Linear Regression, where they modeled the linear dependency of one attribute on others for classification. Logistic Regression was also applied using a sigmoid function, making it suitable for binary classification. Support Vector Machine, another supervised learning algorithm used, performed classification by constructing a hyperplane that best separates the data. Among all the techniques and validation methods, the highest accuracy of 83.83% was achieved using Logistic Regression with 5-fold cross-validation. Similarly, the best result from 10-fold validation was 83.82%, showing consistent and reliable performance across different models and validation strategies.

3. Dataset

We conducted a computer simulation using the heart disease dataset available from the UCI Machine Learning Repository [10]. This dataset includes 303 samples, each with 14 input features and 1 output feature. The input features provide medical and demographic details relevant to heart disease prediction. The output feature represents the presence (value 1) or absence (value 0) of heart disease. However, the paragraph appears to contain some information mistakenly related to a credit dataset, including references to "loan applicants," "Good credit," and "Bad credit." If you intended to describe the heart disease dataset, these references should be removed. If this was meant to refer to a different credit-related dataset, please clarify.

Table-1 List of Features

S.No	Feature Name	Feature Code	Description
1	Age	Age	Age In Year
2	Sex	Sex	Male-1 Feamle-0
3	Type of chest pain	CPT	1-atypical angina 2-Typical angina 3-asymptomatic 4-nonanginal pain

4	Resting Blood Pressure	RBP	mm Hg Admitted as the Hospital
5	serum cholesterol	SCH	in mg/dl
6	Fasting Blood sugar>120 mg/dl	FBS	Fasting Blood sugar>120 mg/dl (1 True,0 False)
7	Resting Electrocardiographic result	RES	0-Normal 1-having ST-T 2-hypertrophy
8	Maximum heart rate achieved	MHR	-
9	Exercise-indued angina	EIA	1-Yes 0-No
10	Old Peak_ST depression indued by Exercise relative Rest	OPK	-

4. Proposed System

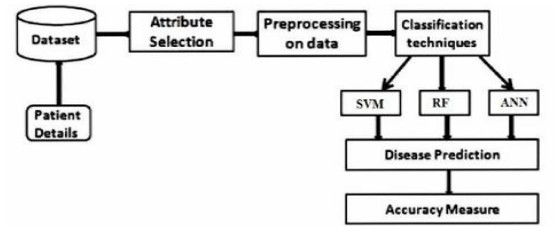

Figure1: Proposed System

4.1 Random Forest

Random Forest is a supervised machine learning algorithm that can be applied to both regression and classification tasks, though it typically performs better in classification. As the name implies, Random Forest involves an ensemble of decision trees. The technique combines multiple decision trees to make a final prediction, based on the principle that the larger the number of trees, the more likely the model will make accurate decisions. For classification tasks, it uses a voting system to determine the most frequent class, while for regression, it averages the outputs of all the decision trees. Random Forest is particularly effective with large datasets that have high dimensionality.

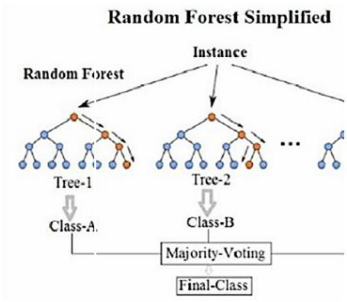

Figure 2: Random Forest

4.2 Support Vector Machines (SVMs)

Support Vector Machines (SVM) come in both linear and non-linear forms. SVM is a supervised learning algorithm commonly used in classification tasks. Typically, two datasets are involved: a training set and a test set. In an ideal scenario, the classes are linearly separable, meaning there exists a line that can perfectly divide the two classes. However, multiple lines can achieve this separation, and the best one is selected as the "separating line."

To prevent overfitting, an SVM allows for some errors, minimizing the number of mistakes made while still achieving a good separation. SVM classifiers have been applied in many fields due to their strong empirical performance. When compared to other classifiers like Naive Bayes, SVMs generally perform better in many applications, making them highly popular in recent research.

4.3 Artificial Neural Network

Neural Networks (ANN) are used to model and simulate the relationships, functions, or mappings among variables within a dynamic system. These networks consist of modules that represent neurons, similar to those in the human nervous system. These neurons are connected through synapses, simulating the interconnections between them.

A Neural Network is built by stacking multiple neurons in layers to produce an output. The first layer is the input layer, and the last one is the output layer. The layers between the input and output layers are called hidden layers. Each neuron in these layers has an activation function. Some common activation functions include Sigmoid, ReLU, and Tanh.

The network's parameters include the weights and biases of each layer. The goal of a neural network is to adjust these parameters so that the predicted output matches the ground truth. Backpropagation, combined with a loss function, is used to optimize the network parameters during training, ensuring the predicted outcomes align with the desired results.

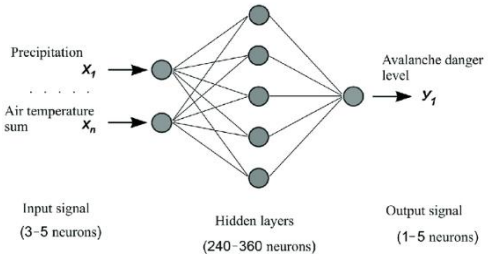

Figure 3: Support Vector Machines (SVMs)

5. Software used

5.1 Python

A web scraper developed using Python was utilized for data collection. Python's concise and readable syntax allows developers to write efficient code with fewer lines, making it ideal for such tasks. The language was originally created by Guido van Rossum at the Centrum Wiskunde & Informatica (CWI) in the Netherlands in

December 1989. Subsequent major versions include Python 2.0, released on October 16, 2000, and Python 3.0, released on December 3, 2008.

Python is a popular choice for web scraping because of its simplicity and the wide range of libraries it offers. Modules like urllib2 (in Python 2) or urllib.request (in Python 3) make it easy to access and extract data from websites. In this study, Python was used to build a scraper that gathered weather data needed for the model.

5.2 Excel

Microsoft Excel is a powerful spreadsheet application developed by Microsoft for both Windows and Mac OS X platforms. It offers a wide range of features including complex calculations, graphing tools, pivot tables, and support for macro programming through Visual Basic for Applications (VBA). The first version of Excel was released in 1987.

Why choose Excel over other spreadsheet software? Excel is a highly versatile tool that supports almost any file format and offers extensive functionality. Its user-friendly interface makes it accessible for beginners, while its advanced features cater to experienced users. Beyond typical tasks like creating charts or performing calculations, Excel allows the creation of custom functions using VBA. It can even be used similarly to an SQL database, providing robust data manipulation capabilities.

Throughout this project, Microsoft Excel has been extensively used for data visualization and preprocessing tasks such as data cleaning, making it an essential tool in the workflow.

6. Result and Discussion

This project focuses on determining whether a patient has heart disease or not [15]. The dataset used was divided into training and testing sets after undergoing preprocessing. Various data classification techniques were applied, including Support Vector Machine (SVM), Artificial Neural Network (ANN), and Random Forest. The project involved a detailed analysis of the heart disease datas*et al*ong with appropriate data processing. Three models were developed, trained, and evaluated, achieving the following maximum accuracy scores:

1) **Support Vector Classifier** – 84.7%
2) **Neural Network** – 83.9%
3) **Random Forest Classifier** – 81.0%

Conclusion

This project offers a comprehensive understanding of machine learning techniques applied to the classification of heart diseases. Classifiers play a vital role in the healthcare sector, as their results can support accurate diagnosis and inform treatment strategies for patients. Existing classification methods have been reviewed and compared to identify the most efficient and accurate models. Machine learning techniques have demonstrated a significant improvement in predicting cardiovascular risk, enabling early detection of heart conditions and facilitating timely preventive care. The findings highlight the vast potential of machine learning algorithms in the prediction of heart-related diseases. However, while each algorithm shows strong performance in certain cases, their effectiveness may vary depending on the specific dataset or scenario.

References

[1]. Mr. ChalaBeyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Technique", International Journal of Pure and Applied Mathematics, 2018.

[2]. Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava, "Effective heart disease prediction using hybrid machine learning techniques" IEEE Access 7 (2019): 81542-81554.

[3]. Ali, Liaqat, *et al*, "An optimized stacked support vector machines based expert system for the effective prediction of heart failure" IEEE Access 7 (2019): 54007-54014.

[4]. Singh Yeshvendra K., Nikhil Sinha, and Sanjay K. Singh, "Heart Disease Prediction System Using Random Forest", International Conference on Advances in Computing and Data Sciences. Springer, Singapore, 2016.

[5]. Prerana T H M1, Shivaprakash N C2 , Swetha N3 "Prediction of Heart Disease Using Machine Learning ,Algorithms- Naïve Bayes, Introduction to PAC Algorithm, Comparison of Algorithms and HDPS" International Journal of Science and Engineering Volume 3, Number 2 – 2015 PP: 90-99

[6]. B.L DeekshatuluaPriti Chandra "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm" International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013.

[7]. Michael W.Berryet.al, Lecture notes in data mining, World Scientific(2006)

[8]. S. Shilaskar and A.Ghatol, "Feature selection for medical diagnosis :Evaluation for cardiovascular diseases," Expert Syst. Appl., vol. 40, no. 10, pp. 4146–4153, Aug. 2013.

[9]. C.-L. Chang and C.-H. Chen, "Applying decision tree and neural network to increase quality of dermatologic diagnosis," Expert Syst. Appl., vol. 36, no. 2, Part 2, pp. 4035–4041, Mar. 2009.

[10]. T. Azar and S. M. El-Metwally, "Decision tree classifiers for automated medical diagnosis," Neural Comput. Appl., vol. 23, no. 7–8, pp. 2387–2403, Dec. 2013. [10] Y. C. T. Bo Jin, "Support vector machines with genetic fuzzy feature transformation for biomedical data classification.," Inf Sci, vol. 177, no. 2, pp. 476–489, 2007.

[11]. N. Esfandiari, M. R. Babavalian, A.-M. E. Moghadam, and V. K. Tabar, "Knowledge discovery in medicine: Current issue and future trend," Expert Syst. Appl., vol. 41, no. 9, pp. 4434–4463, Jul. 2014.

[12]. E. Hassanien and T. Kim, "Breast cancer MRI diagnosis approach using support vector machine and pulse coupled neural networks," J. Appl. Log., vol. 10, no. 4, pp. 277–284, Dec. 2012.

[13]. Sanjay Kumar Sen 1, Dr. Sujata Dash 21Asst. Professor, Orissa Engineering College, Bhubaneswar, Odisha – India.

[14]. Domingos P and Pazzani M. "Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier", in Proceedings of the 13th Conference on Machine Learning, Bari, Italy, pp 105-112,1996.
[15]. Elkan C. "Naive Bayesian Learning, Technical Report CS97-557", Department of Computer Science and Engineering, University of California, San Diego, USA, 1997.
[16]. B.L Deekshatulua Priti Chandra "Reader, PG Dept. Of Computer Application North Orissa University, Baripada, Odisha – India. Empirical Evaluation of Classifiers Performance Using Data Mining Algorithm" A review on genetic algorithm models for hadoop mapreduce in big data
[17]Chandra Shekhar Gautam, Pandey (2019) International Journal of Recent Scientific Research ISSN:0976-3031, Vol.13, Issue-03(E), Pageno771-775, June 2022
[18] Clustering of Bigdata Using Genetic Algorithm in Hadoop MapReduce Chandra Shekhar Gautam, Mr. L N SONI, P Pandey European chemical bulletin Year 2022, issue 12,963-973
[19]Predicting heart disease using machine learning classification technique, Miss Sanjana Chaudhary,Chandra Shekhar Gautam1 and Dr. Akhilesh A Waoo2,Journal of the Maharaja Sayajirao University of Baroda,Volume 10,Year 2024.