Predictive Modelling of Cardiovascular Diseases Using Advanced Algorithms

V. Sowmya Devi[a*], Nikitha Kukunuru[b], Ch. Niranjan Kumar[a], P. Punitha[c]
[a] Department of CSE, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India
[b] Lecturer, Department of Computer Science, MJPTBCWRDC for Women, Siddipet, Telangana, India
[c] Department of CSE(DS), Vignana Bharathi Institute of Technology, Hyderabad, Telangana, India

Abstract: In our day-to-day lives, we have seen many patients suffering from heart disease. Most heart disease diagnoses are based on a complicated combination of clinical and pathological data. Because of this complication, clinical practitioners and researchers are keenly interested in the efficient and precise prediction of cardiac disease. In this paper, we ascertain whether the individual is probable to receive a heart disease diagnosis or not. This work presents machine learning algorithms applied to the real-world dataset from Kaggle. The paper demonstrated three classification methods such as Logistic Regression (LR), Naïve Bayes (NB), and K-Nearest Neighbours (KNN), to build the prediction. The models underwent training using a dataset that was divided into an 80:20 ratio. This provides us with valuable insights that can aid in the forecasting of the number of individuals affected by cardiovascular conditions. The implementation is conducted via Google Colab.

Keywords: Logistic Regression, KNN, Naive Bayes, Histogram, correlation matrix, and cardiovascular disease.

1. Introduction

Machine learning (ML) is “field of study that gives computers the ability to learn without being explicitly programmed” [1]. It is a branch of study devoted to comprehending and developing methods that "learn" - that is, approaches that use data to enhance performance on a set of tasks [2]. Cardiovascular Disease (CVD) is another term for heart disease. It's the main reason people die around the world. The World Health Organisation estimates that CVDs cause 17.9 million deaths worldwide [3]. Machine learning employs a variety of classifiers from Supervised, Unsupervised, and Ensemble Learning to predict and determine the correctness of a given dataset. By analysing patient data and applying a machine learning algorithm to categorise patients as having heart disease or not, this paper aims to predict the future heart ailment. We employ numerous factors to learn about the patient's medical history, such as age, gender, kind of chest pain, serum cholesterol, resting blood pressure, fasting blood sugar, resting electrocardiographic data, the maximum heart rate obtained, exercise induced angina, and old-peak. The purpose of this research is to identify patients who, given their medical history, are at high risk for developing cardiovascular heart problems. A dataset with patients' medical records and other identifying information are chosen from the Kaggle repository for this purpose.

1.1 Motivation

The main goal of this research is to create a heart disease prediction model that can accurately forecast the likelihood of heart disease occurrence. This research endeavour is focused on determining the optimal classification method for assessing the likelihood of heart disease in individuals. The justification for this study is based on doing a comparison analysis utilising three classification algorithms: Naïve Bayes, K-NN, and Logistic regression. The utilisation of confusion matrix and correlation varies throughout different levels of examination. The task of predicting cardiac disease is of utmost importance, requiring the highest level of accuracy, despite the widespread use of these machine learning techniques. Therefore, the algorithms are assessed using various kinds of evaluation procedures. This will empower researchers and medical experts to enhance their predictive capabilities.

1.2 Contribution

This research presents an analysis of machine learning techniques, includes KNN, LR, and NB with the aim of aiding practitioners and medical analysts in accurately diagnosing cardiovascular disease. This study involves an analysis of the latest journals, published works, and data pertaining to cardiovascular illness. This provides a holistic understanding, fostering advancements that can potentially refine and optimize the diagnosis and prognosis of cardiovascular conditions, thereby significantly benefiting the medical community and patients alike.

1.3 Related Work

Various methodologies for predicting heart disease have been proposed in recent years. The utilisation of several ensemble classifiers has been shown to significantly improve the accuracy of heart disease risk prediction, with a reported accuracy rate of 85.4% [4]. The system collects input from the Internet of Things (IoT) devices and securely saves it in a cloud-based storage infrastructure. Based on the current and comprehensive medical records of the patient, the accuracy of forecasting the possibility of heart disease using Bi-LSTM is reported to be 98.86% [5]. Cardiovascular disease diagnosis and prognosis are critical medical duties to ensure precise classification, which allows cardiologists to deliver appropriate treatment to patients. Machine learning applications in the medical field have grown in popularity because they can recognize patterns in data. The authors introduced a k-mode clustering strategy that makes use of Huang starting points to boost classification precision. Experiments with XGBoost, a random forest, a decision tree, and multilayer perceptron models are employed. It had an accuracy of 87.28% [6]. Using Blockchain and ML, numerous parameters such as blood pressure, cholesterol, the blood sugar level, heartbeat count, and body weight may be monitored in order to forecast heart disease at an early stage. It collects data while predicting cardiac disease. SVM achieves 98.2% accuracy [7]. [8] KNN has an accuracy of 88.52%. Models for predicting cardiac disease are provided. [9] [10] [11]. [12] Proposed a cardiac disease model that employs machine learning and deep learning methods.

2. Methodology

Methodology provides a structure for the suggested system model [13]. The technique is a procedure that includes stages that convert given input data into recognised patterns for users' knowledge. The first phase of the proposed work is the gathering of data, the second stage is the extraction of significant values, and the third stage is the preprocessing where we explore the data. The pre-processed data is then split into training and test data and training is done using the machine learning models and then testing is done for classification. The following Fig. 1 shows the model for the proposed system.

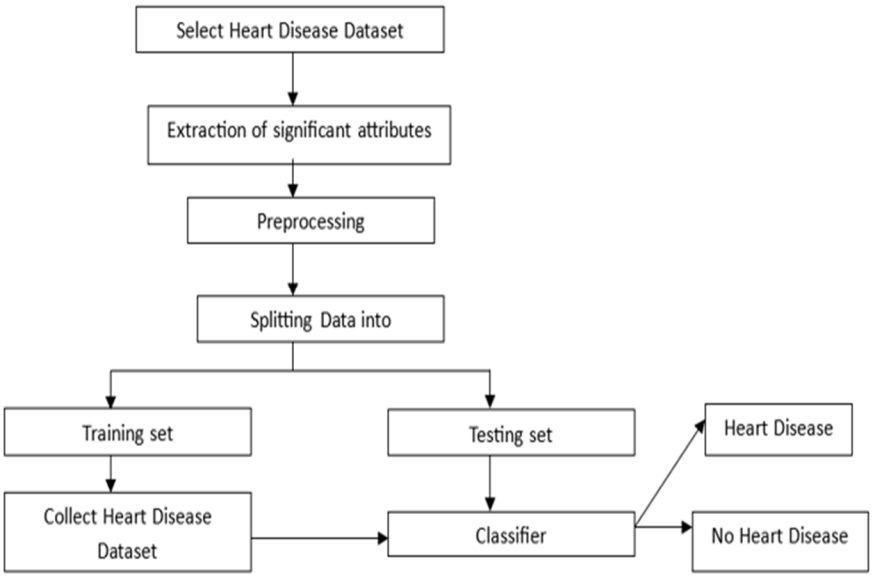

Fig. 1. Proposed System Model

2.1 Dataset

The dataset contains 1205 unique data points. The dataset contains 14 columns [14]. The data set is split into rule of 80%-20%. Where 80% is training and 20% is test data. There are 526 samples with heart disease and 499 with no illness. The following Table 1 shows the 10 records first 10 records of the data set.

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212	0	1	168	0	1	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
58	0	0	100	248	0	0	122	0	1	1	0	2	1
58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
54	1	0	122	286	0	0	116	1	3.2	1	2	2	0

Table 1. Dataset with sample record

The minimum and maximum values of each attribute are given in Table 2.

Sl. No.	Attribute	Description	Minimum and maximum values
1	age	Displays age of the patient	29 to 77
2	sex	Displays the gender of the patient 1=Male, 0=Female	0 and 1
3	cp	Displays the chest pain experienced by the patient (1=angina,2=atypical form of angina,3=non-angina,4=asymptotic angina)	0 to 3
4	trestbps	Displays the resting blood pressure value of an individual in mmHg (unit)	94 to 200
5	chol	Displays the serum cholesterol in mg/dl (unit)	126 to 564
6	fbs	Fasting blood sugar value>120mg/dL (1=true and 0=false)	0 and 1
7	restecg	Value of ECG at rest (0 = normal, 1 = abnormal (ST-T wave), 2 = definite ventricular)	0 to 2
8	thalach	Maximum heart rate recorded	71 to 202
9	oldpeak	Exercise induced ST Depression	0 to 6.2
10	exang	Exercise induced angina(1=yes;0=no)	0 and 1
11	slope	Slope of T segment peak exercise (1=unsloping,2=flat, and 3=down sloping)	0 to 2
12	ca	Major vessels number (0-3) coloured by fluoroscopy	0 to 4
13	thal	3=normal;6=fixed defect;7=reversible defect	0 to 3
14	target	The predicted heart disease status (0=no and 1=yes)	0 and 1

Table 2. Attribute Description

2.2 Data preprocessing

Data preprocessing addresses the aspects cleaning, missing values and normalization [15]. After data has been pre-processed, a classifier is used to divide the data. The proposed method employs K-NN, Logistic Regression, and Naive Bayes as classifiers. Finally, the proposed model is implemented, after which performance and accuracy of our model was evaluated using various metrics. This method employs 14 medical parameters, including cardiac pain, fasting glucose, blood pressure, cholesterol, age, and sex for prediction [16].

2.3 Prediction

a. Logistic regression: It is a statistical technique employed to forecast the result of a categorical dependent variable. Consequently, the result must be a value that falls into a specific category or is discrete in nature. The binary nature of many phenomena allows for categorization into two distinct states, commonly represented as Yes

or No, 0 or 1, True or False, among others. The sigmoid function serves as an activation function. It transforms a predicted real value into a probability value with the range of zero to one.

Sigmoid function:

$$P(y) = 1/(1 + e^{-y}) \quad (1)$$

Eq. (1) shows that P(y) is a probability estimation function, the variable y serves as the input for the probability function, while the mathematical constant e represents Euler's number, which has an approximate value of 2.71828.

Logistic Regression bears the resemblance to Linear Regression, although with differences in their respective applications. Linear regression is commonly employed in the field of statistics and the machine learning to address regression problems, which involve predicting continuous numerical values. On the other hand, logistic regression is frequently utilised to tackle classification challenges, which involve assigning categorical labels to data points. In this study, rather than employing a regression line, we utilise a logistic function with an "S" shape to make predictions. This logistic function is capable of estimating two potential maximum values, either 0 or 1 [17].

b. K-Nearest Neighbours (KNN): K-NN is one of the easiest ML algorithms. It uses a technique called "Supervised Learning". This program looks at how similar the new case or data with the other cases and puts it in the category that is most like the other categories. It saves all the data that was available and sorts a new data point based on how close it is to the old ones. This implies that as new information becomes available, the K-NN algorithm can effectively categorize it into appropriate groups. While the K-NN technique is applicable to both Regression and Classification tasks, its predominant use is in Classification. Operating as a non-parametric method, this refrains from making assumptions about the underlying data. Referred to as a "lazy learner" algorithm, this doesn't swiftly learn from the training set; as an alternative, it retains the dataset and pertains processing when classification is required [18].

Euclidean distance of the two points, p_1 and p_2 is given as Eq. (2).

$$d(c_1, c_2) = \sqrt{(r_1 - r_2)^2 + (q_1 - q_2)^2 + \cdots \ldots + (r_n - r_m)^2} \quad (2)$$

Where:

$d(c_1, c_2)$ is the Euclidean distance b/w points c_1 and c_2.

$(r_1, q_1, ..., r_n)$ are the feature values of point p_1.

$(r_2, q_2, ..., r_m)$ are the feature values of point p_2.

It is computationally expensive for large datasets since it requires calculating distances between the new point and all training points.

KNN serves as a fundamental baseline model in many machine learning tasks and provides a starting point for understanding classification and regression algorithms.

c. Naïve Bayes: A supervised learning technique, which is used to solve problems of classification, the Naive Bayes algorithm is based on Bayes' theorem. This method finds its most prevalent use in text classification tasks involving high-dimensional training datasets. Serving as one of the simplest and yet highly efficient algorithms for Classification, it enables the creation of fast and anticipatory machine learning models. Operating as a probabilistic classifier, it generates predictions grounded in the probability of an object's attributes. The Naive Bayes Algorithm is commonly employed in tasks such as spam filtering and sentiment analysis [19].

The likelihood term in Bayes' theorem is the product of individual conditional probabilities for each feature as shown in Eq. (3)

$$P(Data|Class) \approx P(Z_1|Class) * P(Z_2|Class) * \ldots * P(Z_n|Class) \quad (3)$$

Naive Bayes is fast, simple, and works well with high-dimensional data. However, its assumption of feature independence might not hold in many real-world scenarios. Despite this simplification, Naive Bayes often performs remarkably well, especially in text classification tasks.

2.3 Histogram

The easiest approach to acquire an overview of the distribution of each attribute in a dataset is histograms. This deals with dividing the data into bins. It tells us how many observations fit inside each visualisation bin. The graphing tool is widely used. Data that is measured on an interval scale can be summarized using this method. It is commonly employed to show the salient characteristics of the data distribution in a manageable format [20]. The Fig. 2 gives the histograms of the attributes considered in the models.

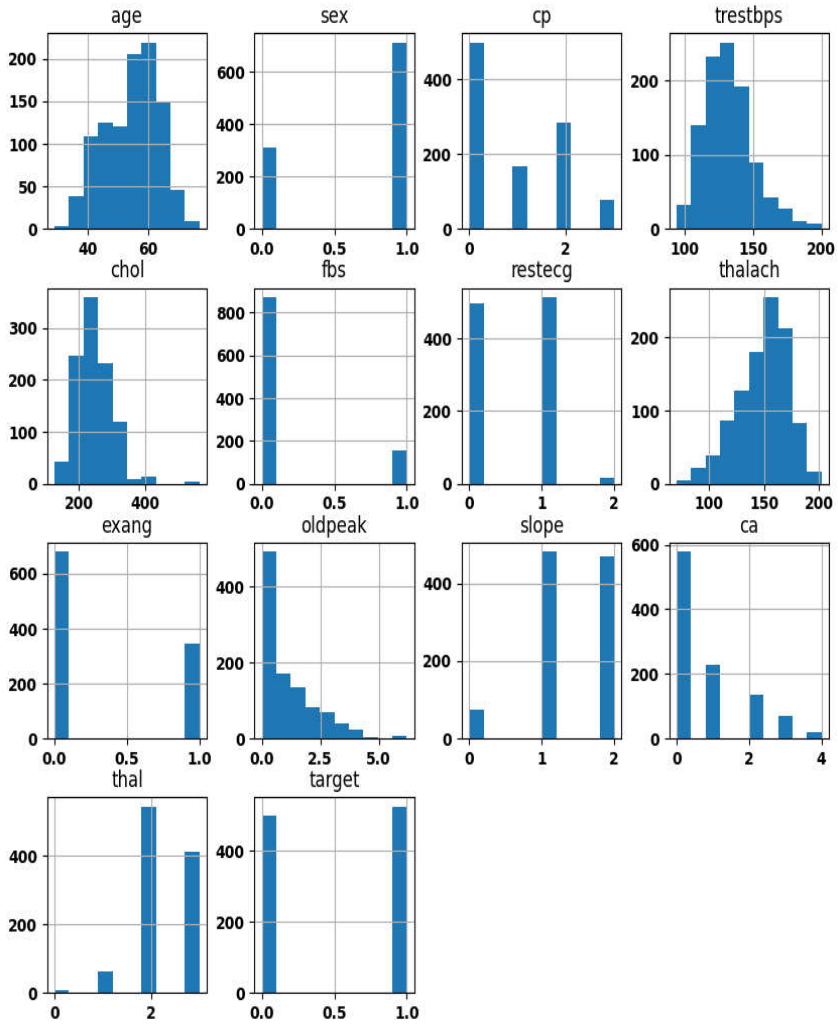

Fig. 2. Histograms of the attributes

2.4 Correlation matrix

It is a table structure that shows the values of the variables' respective correlation coefficients. All potential pairs of data in a table are represented here as a correlation matrix. Summarizing enormous datasets, finding trends, and visualising the results are all possible with this tool [21]. The Fig 3 shows the correlation matrix for the attributes taken for prediction.

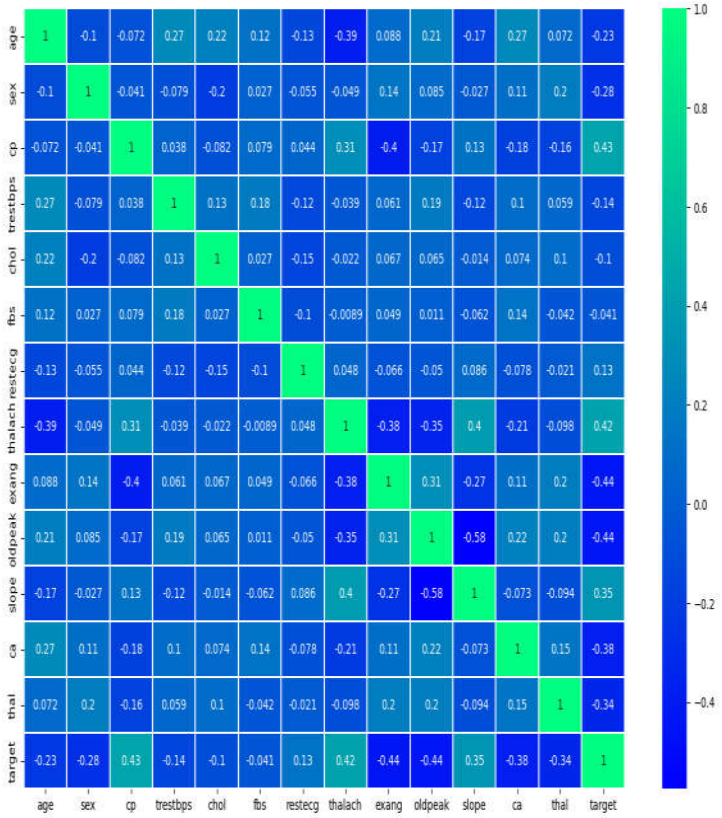

Fig. 3. Correlation Matrix

3. Experimental Setup and performance assessment

The experiment was created and written using Python in Google Colab. SciPy, NumPy, pandas, and Sklearn are just a few of the dependencies that have been implemented as Python modules to make it simple to assess, contrast, and analyse the performance of ML models.

3.1 Evaluation Indices

The K-fold cross-validation technique is a popularly adopted method for evaluating the performance of ML algorithms using a heart disease dataset. Although 'k' often takes on the value of five, which is considered suitable for our heart disease dataset, the algorithms are under scrutiny to investigate how different 'k' values impact the model's estimated performance. The existence of a substantial correlation affirms the reliability of the chosen configuration for the ideal testing situation.
Performance measures like precision, accuracy, specificity, F1 score and recall, are used to decide which machine learning models to use. By comparing the desired output to the real output, accuracy shows how well the model can predict the future. A true negative (TN) and a true positive (TP) show how well the predictor model can forecast whether a patient has cardiovascular or not. The false negative (FN) and the false positive (FP) show that the models made a wrong guess. The precision shows how many real positive observations there are out of all positive cases. Recall figures out how many times the things went well overall, while precision figures out how many times the things went wrong. The F1 score shows the average of both memory and accuracy [8].

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (7)$$

$$Specificity = \frac{TN}{TN + FP} \quad (8)$$

3.2 Results

The Table 3 displays, performance metrics of the three distinct machine learning models.

Table 3. Results using the Machine learning models

Model	Accuracy	Precision	Recall	F1_Score	Specificity
Logistic Regression	0.79	0.76	0.86	0.81	0.83
KNN	0.78	0.76	0.83	0.80	0.81
Naïve Bayes	0.83	0.81	0.86	0.84	0.85

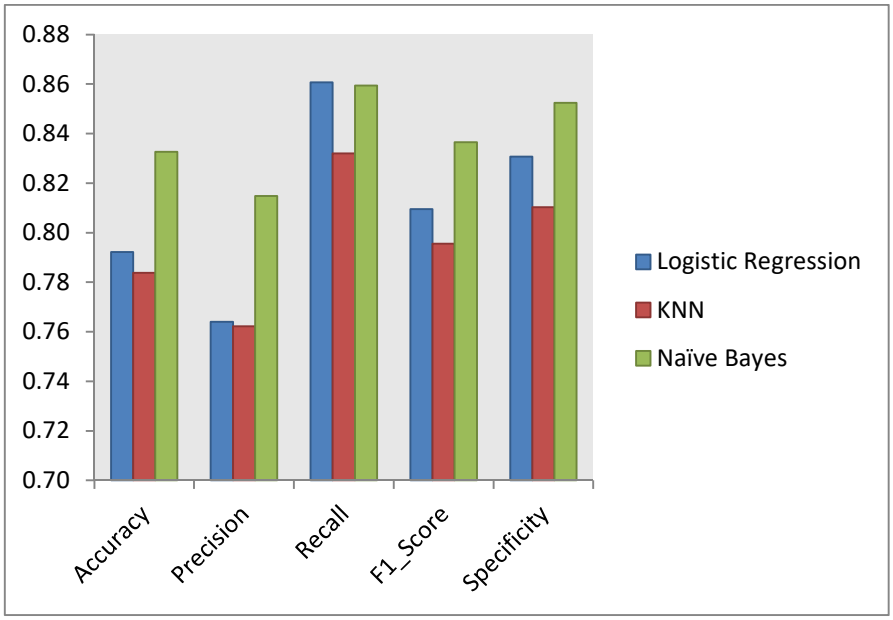

Fig. 4. Analysis of the machine learning models

1. *Accuracy:* This statistic shows the correctness of the model's predictions altogether, or the proportion of cases correctly classified out of every occurrence. Naïve Bayes achieves the best accuracy with a value of 0.83, next by K-NN with 0.78 and Logistic Regression with 0.79.

2. *Precision*: This is a metric that expresses the percentage of correctly predicted positive cases among all the model's positive predictions. This illustrates how well the algorithm can reduce false positives. The Naïve Bayes algorithm has the highest level of precision, with a score of 0.81. The K-NN and Logistic Regression algorithms closely follow, both achieving a precision score of 0.76.

3. *Recall*: This is also known as sensitivity or true positive rate, is the proportion of correctly predicted positive occurrences out of all actual positive occurrences. It shows how well the approach can capture good examples. Naïve Bayes and Logistic Regression achieve the highest recall values (0.86), while K-NN has a slightly lower recall of 0.83.

4. *F1 Score:* This is the mean of your recall and knowledge of a subject. It accounts for both false negatives and false positives. With an F1 score of 0.84, Naïve Bayes has the highest score, closely followed by Logistic Regression (0.81) and KNN (0.80).

5. *Specificity:* Out of all actual negative events, specificity is the number of accurate negative forecasts. This indicates the approach's ability to identify instances of mishandling. Naive Bayes approach (0.85), followed by Logistic Regression (0.83) and KNN (0.81), has the highest specificity.

4. Conclusion and Future Scope

Based on the performance metrics, Naïve Bayes seems to outperform both Logistic Regression and K-NN in most aspects. This attains the highest accuracy; recall, precision, and the F1 score among three models, making it a strong choice for this particular classification task.

Logistic Regression and KNN have comparable results, but overall, Naïve Bayes demonstrates a slightly better balance between precision and recall, as indicated by the higher F1 score. If precision and recall are both important for the application, Naïve Bayes is the recommended model.

However, it's essential to consider other factors such as model complexity, interpretability, and computational efficiency when selecting the final model. Based on the specific use case and requirements, any of the three models might be a reasonable choice.

It's worth noting that these results are based on a specific dataset and evaluation procedure. It's essential to validate the models on different datasets, perform cross-validation, and conduct further analysis to ensure the generalizability and reliability of the chosen model. Additionally, hyperparameter tuning and feature engineering could potentially improve the performance of all models.

References

1. Artificial Intelligence in Design '96. Springer, Dordrecht. pp. 151–170. ISBN 978-94-010-6610-5. (1996). doi:10.1007/978-94-009-0279-4_9.
2. Mitchell, Tom, Machine Learning. New York: McGraw Hill. ISBN 0-07-042807-7. OCLC 36417892. Archived from the original on 2020-04-07. Retrieved 2020-04-09. (1997).
3. Dangare C S & Apte S S, " Improved study of heart disease prediction systems using data mining classification techniques," International Journal of Computer Applications, 47(10), 44-8 (2012).
4. Latha, C.; Jeeva, S. "Improving the Accuracy of Prediction of Heart Disease Risk Based on Ensemble Classification Techniques," Inform. Med. Unlocked , 16, 100203 (2019).
5. Nancy, A.A.; Ravindran, D.; Raj Vincent, P.M.D.; Srinivasan, K.; Gutierrez Reina, D. "IoT-Cloud-Based Smart Healthcare Monitoring System for Heart Disease Prediction via Deep Learning," Electronics , 11,2292.(2022). https://doi.org/10.3390/electronics11152292
6. Bhatt, C.M.; Patel, P.; Ghetia, T.; Mazzeo, P.L. Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms, 16, *88*.(2023).
https://doi.org/10.3390/a16020088
7. A. Nouman and S. Muneer, "A Systematic Literature Review On Heart Disease Prediction Using Blockchain And Machine Learning Techniques", IJCIS, vol. 1, no. 4, pp. 1–6, Dec. (2022).

8. Harshit Jindal et al, “Heart disease prediction using machine learning algorithms,” *IOP* Conf. Ser.: Mater. Sci. *Eng.* 1022 012072*, (*2021). doi: 10.1088/1757-899X/1022/1/012072

9. Diaa s AbdElminaam et.al., “ML HeartDisPrediction: Heart disease prediction using machine learning,” JOCC, article 6, Vol.2, Jan (2023). doi: 10.21608/JOCC.2023.282098

10. T. M. Ghazal, A. Ibrahim, A. S. Akram, Z. H. Qaisar, S. Munir and S. Islam, "Heart Disease Prediction Using Machine Learning," International Conference on Business Analytics for Technology and Security (ICBATS)*,* Dubai, United Arab Emirates, pp. 1-6, (2023), doi: 10.1109/ICBATS57792.2023.10111368.

11. C. Boukhatem, H. Y. Youssef and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," Advances in Science and Engineering Technology International Conferences (ASET*)*, Dubai, United Arab Emirates, pp. 1-6, (2022). doi: 10.1109/ASET53988.2022.9734880.

12. K. Vayadande et al*., "Heart Disease Prediction using Machine Learning and Deep Learning Algorithms,"* International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*,* Greater Noida, India, pp. 393-401, (2022). doi: 10.1109/CISES54857.2022.9844406.

13. Wolgast G, Ehrenborg C, Israelsson A, Helander J, Johansson E & Manefjord H. Wireless body area network for heart attack detection [Education Corner]. IEEE antennas and propagation magazine, 58(5), 84-92 (2016).

14. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

15. Zhang Y, Fogoros R, Thompson J, Kenknight B H, Pederson M J, Patangay A & Mazar S T, U.S. Patent No. 8,014,863.Washington, DC: U.S. Patent and Trademark Office. (2011).

16. Buechler K F & McPherson P H U.S. Patent No. 5,947,124. Washington, DC: U.S. Patent and Trademark Office. (1999).

17. Ambrish G, Bharathi Ganesh, Anitha Ganesh, Chetana Srinivas, Dhanraj, Kiran Mensinkal, Logistic regression technique for prediction of cardiovascular disease, Global Transitions Proceedings, vol. 3, no. 1, pp 127-130, ISSN 2666-285X,(2022,).https://doi.org/10.1016/j.gltp.2022.04.008.

18. G. Kaur and A. Chhabra, “Improved J48 classification algorithm for the prediction of diabetes,” International Journal of Computers and Applications, vol. 98, no. 22, pp. 13–17, (2014).

19. G. I. Webb, J. R. Boughton, and Z. Wang, “Not so naive bayes: Aggregating one dependence estimators," Mach. Learn., vol. 58, no. 1, pp. 5-24, Jan. (2005). https://doi.org/10.1007/s10994-005-4258-6

20. Victor Chang, Vallabhanent Rupa Bhavani, Ariel Qianwen Xu, MA Hossain, An artificial intelligence model for heart disease detection using machine learning algorithms, Healthcare Analytics, vol. 2, 100016,ISSN 2772 4425, (2022). https://doi.org/10.1016/j.health.2022.100016.

21. Sumaira Ahmed, Salahuddin Shaikh, Farwa Ikram, Muhammad Fayaz, Hathal Salamah Alwageed, Faheem Khan, Fawwad Hassan Jaskani, "Prediction of Cardiovascular Disease on Self-Augmented Datasets of Heart Patients Using Multiple Machine Learning Models", Journal of Sensors, vol. Article ID 3730303, 21 pages, (2022). https://doi.org/10.1155/2022/3730303.