International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 07 | July - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Heart Disease Prediction Using Machine Learning Algorithms

Mahammad Sahil Khan
Dept. of Computer Science and Engineering
Email ID: mkhan2022@gift.edu.in

Asst.Prof. Archana Panda
Dept. of Computer Science and Engineering
Email ID: archana.panda@gift.edu.in

Abstract- Heart disease is a major issue that has become increasingly prevalent. According to current statistics, heart disease claims the life of one person every minute. In the last several years, one of the hardest problems facing the medical field is predicting heart disease. Reducing the death rate can be achieved with early detection of cardiac disease. Machine learning is the most effective approach to forecasting heart disease. This paper aims to create a lightweight, straightforward solution to detecting cardiac disease using machine learning. Machine learning can aid in heart disease prediction. This study analyzes several machine learning algorithms and performance indicators. This study compares cardiac disease detection methods using a publicly available dataset from the UCI machine learning repository. There are other datasets accessible, including the Switzerland and Cleveland databases. Here the dataset contains 303 patient records and 18 characteristics. The analysis shows that out of six machine learning algorithms, the Random Forest algorithm gives the best result with 94.50%.

Keywords- cardiac disease detection, datasets, heart disease prediction, Machine Learning, Random Forest algorithm.

I. INTRODUCTION

Many severe issues have been brought about by heart disease; one of the main obstacles in this regard is accurately detecting and identifying the disease's presence inside an individual. While there are a number of medical devices on the market that can be used to forecast heart disease, they are very costly and insufficiently accurate to determine the likelihood of heart disease. Improved and more effective methods for early cardiac illness diagnosis are required [1]. This has been made possible by the growth of computer science in various scientific fields, including the medical sciences. As opposed to being explicitly designed, a machine-learning system is learned. A more accurate option for obtaining high detection accuracy for cardiac problems might be machine learning. The purpose of this paper is to provide a broad overview of machine learning techniques used in heart disease prediction.

This paper will go into detail in a later section on different machine learning algorithms and how accurate they are in comparison.

Machine learning algorithms

"Computers can learn and behave like humans with the aid of machine learning algorithms, and their learning can be enhanced by providing them with data and knowledge in the form of observations. Algorithms for machine learning are those that can identify hidden patterns in data, forecast results, and enhance performance via independent experience." There are several machine learning methods out there. Six distinct algorithms were compared and analyzed in this paper.

A. KNN

A supervised learning technique that can be applied to regression and classification issues is the K-Nearest Neighbor. This method works by assuming similarities between the new data point and the existing data points. The new data points are grouped into the most similar groups based on these commonalities.

B. Naive Bayes

It is an algorithm that discovers every object's probability, characteristics, and groupings. Under supervised learning, the Naive Bayes Algorithm is primarily utilized for resolving

classification issues. The probability results that this algorithm can provide for issues that cannot be solved through prediction are the basis upon which it is constructed.

C. SVM

The support vector machine (SVM) is a machine learning algorithm that determines boundaries between data points based on predefined classes, labels, or outputs. The SVM algorithm's main goal, technically, is to locate a hyperplane that clearly divides the data points into separate classes.

D. Decision Tree

A decision tree is a structure that resembles a flowchart and is used to forecast or make decisions. It is made up of nodes that represent attribute tests or decisions, branches that show the results of these tests or decisions, and leaf nodes that show the conclusions or predictions that are reached.

E. Random Forest

A supervised classification algorithm called Random Forest uses many decision trees on different dataset subsets and averages them to increase the dataset's predicted accuracy. Rather than depending on a single decision tree, the random forest forecasts the outcome based on the majority vote of projections from each tree.

F. Logistic Regression

One of the most widely used machine learning algorithms, under the category of supervised learning, is logistic regression. With a given collection of independent factors, it is used to predict the categorical dependent variable. With logistic regression, the result of a categorical dependent variable is predicted. As a result, a discrete or category value must be the result. It provides probabilistic values, which range from 0 to 1.

II. PROPOSED MODEL

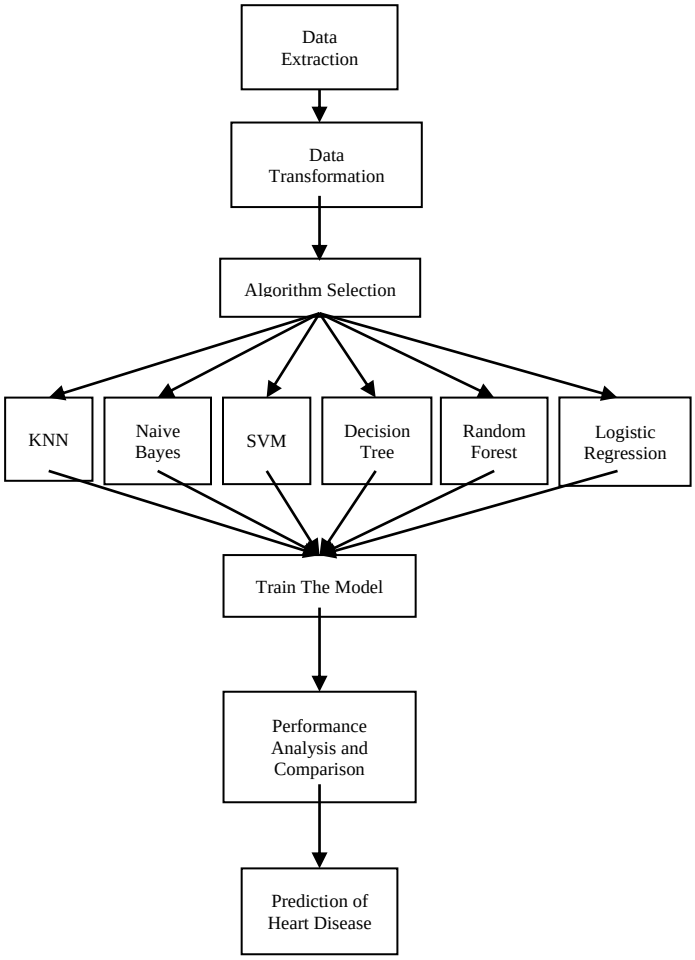

Figure 1: Proposed work

Figure 1 shows the proposed model. Where the first step is Data Extraction, Collecting the dataset for analysis and comparison from multiple sources, the second step is Data Transformation. Data Transformation is a process of removing all the outliers, missing values, converting and structuring into usable format. The third step is Algorithm Selection where different algorithms are selected and used in model building. In the fourth step models are trained from the past data. In the next step, performance is analyzed and the accuracy of the algorithms is compared and find the best algorithm based on accuracy. At last, it will predict whether a person has heart disease or not based on the given input.

III. METHODOLOGY

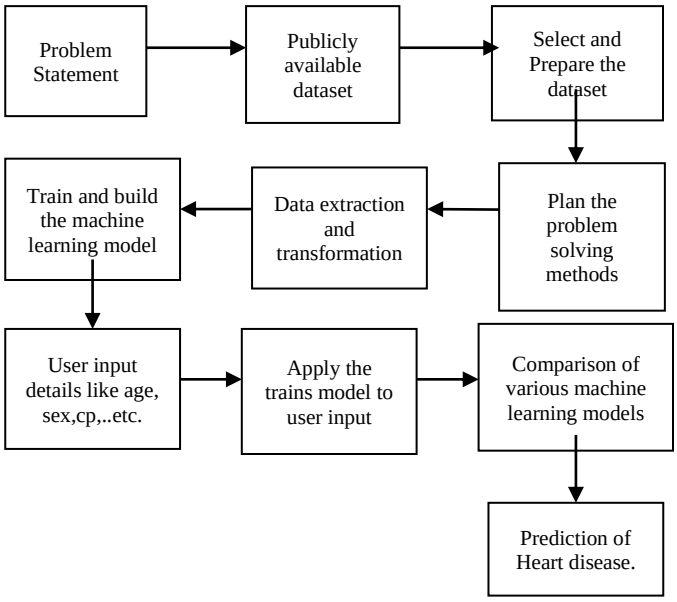

Figure 2: Design and Approach

The project plan is depicted in Figure 2, where the problem is described in the problem statement. The next step is to find the publicly available dataset. The UCI machine learning repository has a large number of datasets, including datasets from Cleveland, Switzerland, and Hungary. The right dataset must be chosen in the following step. Here, Cleveland dataset is used for analysis and comparison in this paper. The following stage involves organizing the problem-solving techniques or approaches employing five classifiers: Random Forest, Naive Bayes, KNN, Decision Tree, and Support Vector Machine, or SVM. Data extraction and transformation are the following steps. Gathering the dataset for analysis and comparison from various sources is known as data extraction, and removing all outliers and missing values as well as converting and organizing the data into a format that can be used for use is known as data transformation. Creating a training model for prediction is the next stage. To do this, use publically available datasets that include information on age, sex, and other characteristics. provide training by the use of several machine learning algorithms, and the outcomes were measured using various performance indicators. The next stage is to forecast heart disease using user data and the built model. Comparing different machine learning algorithms based on classification results is the next stage. It will compare each classifier in the final step and predict the person has heart disease or not.

A. Problem Statement

Examine the Cleveland dataset to determine if a person is at risk for heart disease. A person's heart disease is represented by the number 1, and their absence of heart disease is represented by the number 0.

B. Select Dataset

The dataset is a collection of data that is normally presented in a tabular form. The UCI machine learning repository has a large number of datasets [2]. Among the datasets are those from Cleveland(https://archive.ics.uci.edu/ml/datasets/Heart+Disease), Switzerland, and Hungary. [3]. For this investigation, only 18 attributes are employed. These characteristics are as follows: https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset

1. Age
2. Sex
3. Chest Pain
4. Trestbps
5. Chol
6. Fbs
7. Restecg
8. Thalach
9. Exang
10. Old peak
11. Slope
12. Ca
13. Thallium Heart Rate
14. Alcohol intake
15. Smoking
16. Diet
17. Stress
18. Heart_disease_Predict_lebel

C. Problem Solving Strategy

The following machine learning techniques—decision trees, artificial neural networks, support vector machines, and Naive Bayes—are used to make effective decisions in the healthcare industry. Here, six distinct algorithms—including logistic regression, KNN, naïve bayes, random forest, decision tree, and svm—were employed for comparison.

D. Data extraction and transformation

Gathering the dataset for analysis and comparison from various sources is known as data extraction, and removing all outliers and missing values as well as converting and organizing the data into a format that can be used for use is known as data transformation.

E. Train and build machine learning model for heart disease detection

The dataset is split into two sections in this step: the training dataset and the testing dataset. 60% of the training dataset and 40% of the testing dataset are randomly picked.

F. Input Details

User input details include age, gender, cp, Trestbps, chol, Fbs, Exang, Thalach, old peak, slope, ca, thal, restecg, and class. There are 17 characteristics and 1 label.

G. Comparison of various machine learning algorithms

This stage involves comparing the classifiers.

Various classifiers, including svm, random forest, knn, naive bayes, decision tree, and logistic regression, are compared based on accuracy, precision, recall, and F1 score.

H. Prediction of heart disease

The Python programming language is used to input all necessary values for evaluation. After receiving user input, the trained model determines if a person has heart disease.

IV. RESULTS

Table 1: Values obtained for confusion matrix using different algorithms

	Algorithm	True Positive	False Positive	False Negative	True Negative
1	NB	21	6	3	31
2	SVM	21	5	3	30
3	K-NN	22	5	4	30
4	DT	25	2	4	30
5	RF	22	5	6	28

Correlation Matrix

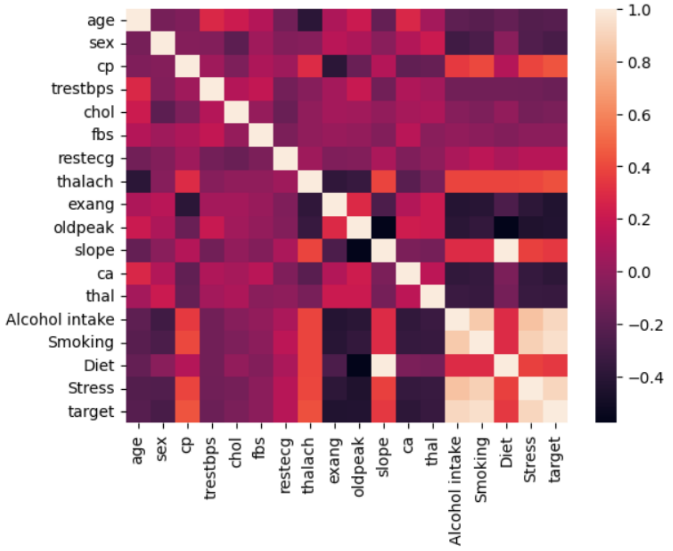

Table 2: Classification results

	Algorithm	Accuracy	Precision	Recall	F-1Score
1	Naive Bayes (NB)	86.7	85.7	91.9	88.2
2	SVM	91.73	92.3	88.2	89.1
3	K-NN	87.11	86.06	84.41	87.2
4	DT	84.98	85.9	89.16	84.3
5	RF	94.50	93.7	89.2	91.16

V. CONCLUSION

The Development of a system that can reliably and efficiently predict heart illness is becoming more and more important due to the increasing number of heart disease deaths. The study sought to identify the most effective machine learning method for detecting heart problems. This study examined the accuracy of DT, KNN, RF, SVM, and NB algorithms for heart disease prediction. It analyzes data such as blood pressure, cholesterol, chest pain, alcohol intake, smoking, diet, and stress to help estimate a patient's risk of a heart disease. The dataset needs to be normalized in order to prevent the training model from overfitting and from producing insufficient accuracy when a model is evaluated for real-world data problems, which can differ greatly from the dataset used for training. It was also discovered that statistical analysis plays a significant role in the analysis of a dataset. More datasets can be used to enable deep learning with a variety of additional improvements, potentially yielding more encouraging outcomes. The data can be normalized in more ways, and the outcomes can be contrasted. Additionally, there may be more ways to combine specific multimedia with ML and DL models trained on cardiac diseases for the convenience of physicians and patients.

FUTURE SCOPE

Machine learning models can forecast disease risk based on patient data. AI makes it possible to personalize treatment regimens according to a patient's genetic composition and medical background. This precision medicine method can lessen side effects and increase therapeutic success. Diagnosticians can decrease misdiagnosis by classifying cardiovascular disease occurrence using machine learning. In order to lessen the death rate from cardiovascular disorders, this study creates a model that can accurately forecast these conditions.

REFERENCES

[1] L. Gao, Y. Ding
Disease prediction via Bayesian hyperparameter optimization and ensemble learning
BMC Res Notes, 13 (2020), pp. 1-6

[2] https://medium.com/analytics-vidhya/heart-disease-prediction-with-ensemble-learning-74d6109beba1.

[3] A.E. Hegazy, M.A. Makhlouf, G.S. El-Tawel
Improved salp swarm algorithm for feature selection
J. King Saud Univ. Comput. Inf. Sci., 32 (3) (2020), pp. 335-344

[4] Y. Khourdifi, M. Bahaj
Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization Int. J. Intell. Eng. Syst., 12 (1) (2019), pp. 242-252

[5] Drożdż, K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman, W.; Nalepa, J.; Gumprecht, J.; Lip, G.Y.H. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. Cardiovasc. Diabetol. 2022, 21, 240.

[6] Arsalan Khan, Moiz Qureshi, Muhammad Daniyal, Kassim Tawiah
A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction First published: 20 February 2023 https://doi.org/10.1155/2023/14060