EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

A Novel Approach to Heart Disease Prediction Using Artificial Intelligence Techniques

V. Sathyavathy[1,*]

[1]Department of Computer Technology, KG College of Arts and Science, India

Abstract

INTRODUCTION: Heart disease remains one of the leading causes of mortality worldwide, necessitating the development of accurate and efficient prediction models
OBJECTIVES: To research new models for heart disease prediction
METHODS: This paper presents a novel approach for predicting heart disease using advanced artificial intelligence (AI) techniques, including machine learning (ML) and deep learning (DL) algorithms
RESULTS By leveraging patient data and integrating various AI models, this approach aims to enhance prediction accuracy and support early diagnosis and intervention
CONCLUSION: This study presents a novel AI-based approach for heart disease prediction, demonstrating the efficacy of ML and DL models in improving diagnostic accuracy

Keywords: Cardiovascular Disease, Random Forest Algorithm, Artificial Intelligence, Logistic Regression

Received on 14 03 2024, accepted on 20 06 2024, published on 30 07 2024

doi: 10.4108/eetpht.10.6807

1. Introduction

The most common cause of death worldwide is heart failure, which is becoming a challenging issue. Globally, the number of instances of heart disease has increased from 271 million in 1990 to 523 million in 2019, and during the same time period, the number of deaths has increased from 12.1 million to 18.6 million. Scalable and reasonably priced solutions are needed to reduce the burden of CVD, and artificial intelligence (AI) may be a crucial element.

2. Role of AI in Health Care

To improve speed and processing capacity, large volumes of data sets may now be perfectly and swiftly examined. By searching for patterns in patient data and helping stroke victims, this has allowed medical professionals to apply AI to large, complicated data sets to improve diagnosis, decision-making, and support.

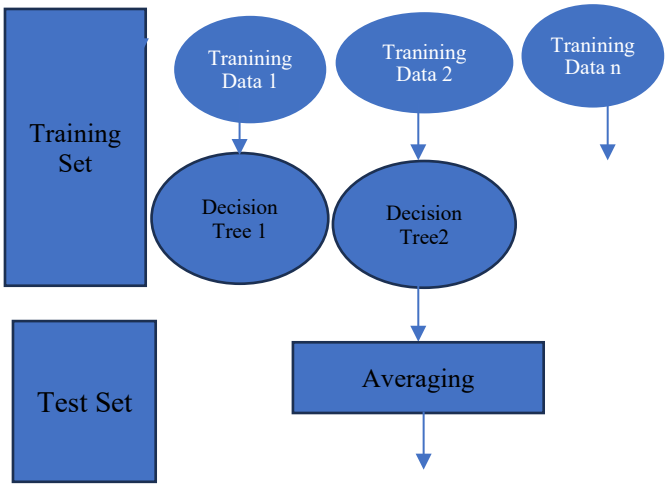

Figure 1. Training Set and Test Set

*Corresponding author. Email: Sathyavathy.v@kgcas.com

2.1. Heart Problems Prevention

AI analysis of ECGs has produced a universally applicable, low-cost diagnostic for detecting the presence of a weak heart pumping blood to the arteries, which can lead to cardiac failure if left untreated. The Mayo Clinic, which has a database containing more than 7 million ECGs, is in a great position to expand this application of AI. To protect privacy, all personally identifiable patient data is initially erased.

Deep learning (DL), a subset of machine learning (ML) that uses artificial neural networks for sophisticated tasks like object recognition and speech analysis, is one of the many types of AI algorithms Applications of deep learning can be found in industries such as healthcare, where it helps with tasks like outcome predicting and diagnostic prediction. In both healthcare research and practice, prediction models are essential. ML prediction models are becoming more popular in cardiology to help with clinical decision-making and outcome forecasting. To help doctors with individualised treatment plans, these models have been used, for example, to forecast mortality after heart transplantation in patients with congenital cardiac disease. Even though AI has a lot of potential applications in healthcare, particularly paediatric cardiology, more research is still required to create interpretable models.

2.2. Prediction of Cardiovascular Disease

AI is rapidly changing several industries, including healthcare, especially when it comes to controlling and detecting CVDs. AI has a wide range of potential uses in the early diagnosis and prognosis of cardiovascular disease. Image capture, optimisation, automated measurements, outlier identification, diagnostic categorization, and result prediction are some of these activities.

2.3. AI and Cardiovascular Disease

Humans and machines can collaborate to optimise and provide healthcare, and each has unique benefits and drawbacks. AI will be developed and used to enhance human intelligence, not to take its place, according to the American Medical Association's most recent description of the technology in healthcare.8 The viewpoint of the American Medical Association emphasises human-machine cooperation, which has important implications for the use of AI in healthcare. Here are our thoughts on how AI works and how to best develop, use, and integrate AI to advance the digital healthcare revolution and enhance human performance.

Supervised Learning

The ultimate goal of supervised machine learning is to create an algorithm that can, for example, predict the occurrence of an outcome or categorise images by using labelled data to build models that establish a relationship between the input data and their associated label.

UnSupervised Learning

Unsupervised machine learning techniques work well for creating models that can find patterns in the input data without the need for labels. This method makes it possible to find connections between features and aids in the discovery of hidden structures (such clusters or patterns) inside a dataset. Typical instances of unsupervised algorithms are hierarchical clustering and k-means. Under unsupervised learning conditions, neural networks (such as convolutional or recurrent ones) are also employed. Convolutional neural networks are particularly noteworthy for their capacity to automatically extract patterns or clusters (including those that are invisible to the human eye) when evaluating complex datasets like transcriptome, proteome, or genomic profiles, as well as MRI, echocardiography, and textual datasets.

2. Machine Learning Algorithms

The following machine learning algorithm is used for comparing the performance metrics
- Logistic Regression
- Decision Trees
- Random Forests
- Support Vector Machines (SVM)
- k-Nearest Neighbors (k-NN)
- Gradient Boosting Machines (GBM)
- Neural Networks

Table 1: Dataset

Attribute	Meaning
Age	Age is Continuous
Gender	1=male 0=female
Cp	Chest Pain
Trestbps	Resting Blood Pressure(mmHg)
Chol	Cholesterol
Fbsl	Fasting blood sugar 0:<=120mg/dl, 1:>120mg/dl
restecg	electrocardiographic results during resting 1=true 0=false
thalach	Maximum heart rate achieved: continuous
exang	Exercise induced angina
oldpeak	ST depression
slope	ST segment slope
ca	Number of major vessels coloured by fluoroscopy:discrete(0,1,2,3)
thal	3: normal 6: fixed defect 7: reversible defect

Random Forest Algorithm

This is the most well-liked and promising algorithm in machine learning. It is a part of machine learning that is supervised. This approach is applied to problems involving both regression and classification. This method collects data, builds decision trees using many data sets, and then averages the decision trees. This method can be used to tackle both regression and classification problems.

Logistic Regression

One machine learning classification method for datasets that examine relationships between the independent factors influencing a result and a classified dependent variable (DV) is logistic regression. The Logistic Regression (LR) Classifier predicts a dependent variable's result using a range of value identifiers. When the dependent component is all out and dichotomous and the autonomous variables are all out, persistent, or integrated, it makes sense.
Two methods of demonstration are possible with logistic regression:
1. Regression stepwise: This is a regression technique in which each autonomous component is added one at a time, and the model's significance is assessed thereafter.
2. Regression in Reverse stepwise: This approach makes use of every independent component that is available and progressively removes irrelevant features to verify that the model correctly matches.

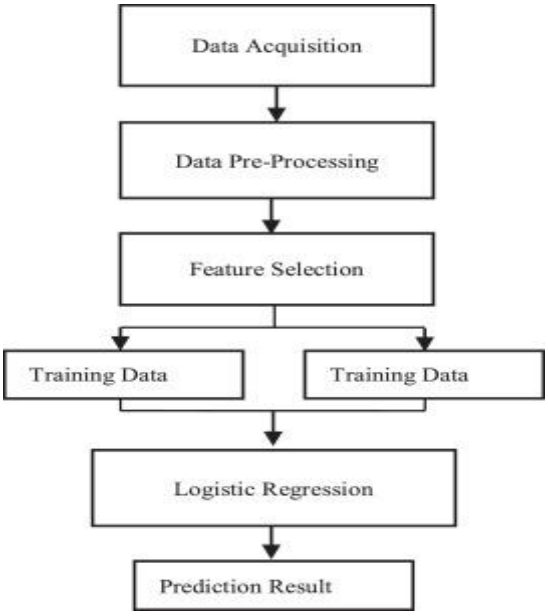

Figure 2. Logistic Regression

Decision Trees

Although decision trees are a supervised learning technique, they are primarily employed to solve classification problems. However, they can also be used to solve regression problems. This classifier is tree-structured, with internal nodes standing in for dataset attributes, branches for decision rules, and leaf nodes for each outcome.
The Decision Node and the Leaf Node are the two nodes that make up a decision tree. While leaf nodes represent the result of decisions and do not have any more branches, decision nodes are used to make any kind of decision and have numerous branches.

KNN Algorithm

One supervised machine learning approach that is frequently used for classification is the k-nearest-neighbor (KNN) algorithm. It is frequently utilised for illness prediction.Using the features and labels of the training data, the supervised KNN algorithm predicts the classification of unlabeled data. By considering the k nearest training data points (neighbours), which are the ones closest to the query it is testing, the KNN algorithm can typically categorise datasets using a training model that is comparable to the testing query.

$$Average\ Accuracy = \sum_{i=1}^{l} \frac{TP_i + TN_i}{TP_i + FN_i + FP_i + TN_i} / l$$

Figure 3. KNN Formula

Support Vector Machine

Heart disease is a major global health concern, and successful treatment and prevention depend heavily on early detection. Based on patient data, machine learning algorithms like Support Vector Machines (SVM) have demonstrated encouraging outcomes in the prediction of cardiac disease.

Support vector machines, or SVMs for short, are a class of supervised learning techniques used in regression analysis, classification, and outlier detection. SVM, also called a hyperplane, is a surface that maximises the boundaries between different kinds of data points that are represented in multi-dimensional space. It is applicable to multi-class classification as well as binary classification.

Neural Networks

Deep learning models, which use neural networks, have demonstrated great promise in the prediction of cardiac disease. Traditional machine learning methods could miss intricate patterns and interactions in the data that these models are able to capture.

A simple multi-layer perceptron (MLP) for heart disease prediction can be designed as follows:

Input Layer: Number of neurons equal to the number of features.
Hidden Layers: One or more hidden layers with a varying number of neurons and activation functions.
Output Layer: A single neuron with sigmoid activation function for binary classification (presence or absence of heart disease).

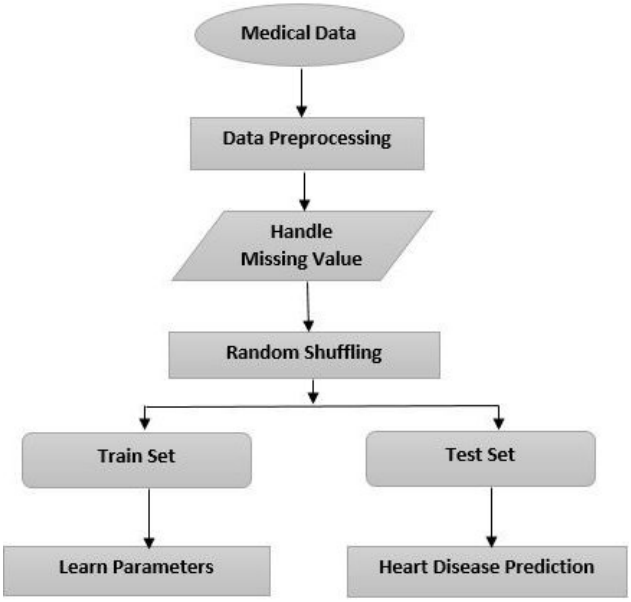

Figure 4. Neural Networks

Gradient Boost Algorithm

Data Collection and Preprocessing:

- Gather heart disease.
- Feature Selection:
- Select relevant features such as age, gender, blood pressure, cholesterol levels, and other clinical measurements.

Model Training:

- Split the data into training and testing sets.
- Train the gradient boosting model using the training set.

Model Tuning:

- Optimize hyperparameters using techniques such as grid search or random search.
- Important hyperparameters include the number of trees, learning rate, maximum depth of trees, and minimum samples per leaf.

Model Evaluation:

- Evaluate the model using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score.

Interpretability:

- Use feature importance scores to understand the impact of each feature on the prediction.

Advantages

(i) High Accuracy: Gradient boosting models generally provide high predictive accuracy due to their ability to minimize bias and variance.
(ii) Handling Various Types of Data: These models can handle numerical and categorical data effectively.
(iii) Feature Importance: Gradient boosting provides insights into feature importance, which helps in understanding the influence of each feature on the predictions.

3.Training Set and Test Set

To assess the model's performance on unobserved data, the dataset must be divided into training and test sets when creating a model for heart disease prediction. This procedure aids in evaluating the model's capacity for generalisation. Here is a comprehensive guidance on how to partition, prepare, and assess the information for the purpose of predicting heart disease.

Evaluating the performance of different machine learning algorithms for heart disease prediction involves assessing their accuracy on both training and test datasets. Here's an example process for training multiple algorithms and comparing their performance.

Preprocessing Steps

- Handle missing values using imputation techniques.
- Normalize numerical features.
- Encode categorical variables.
- Split the data into training and testing sets.

Evaluation Metrics

Accuracy:
Measures the proportion of correctly classified instances.
ROC-AUC Score:
Measures the ability of the model to distinguish between classes. A higher score indicates better performance.

Confusion Matrix:
Provides insights into the true positives, false positives, true negatives, and false negatives.

Train Accuracy: 0.95
Test Accuracy: 0.88
ROC-AUC Score: 0.90
Confusion Matrix:
[[80 10]
[12 48]]

Table 2: Training Set and Test Set

Algorithm	Train Accuracy	Test Accuracy
Logistic Regression	0.85	0.83
Decision Tree	1.00	0.78
Random Forest	0.98	0.86
Support Vector Machine	0.87	0.83
k-Nearest Neighbors	0.85	0.81
Naive Bayes	0.84	0.82
XGBoost	0.97	0.88
LightGBM	0.97	0.88
CatBoost	0.98	0.88

Analysis

Logistic Regression: Moderate accuracy on both training and test sets, indicating good generalization.
Decision Tree: Perfect accuracy on training set but lower on test set, suggesting overfitting.
Random Forest: High accuracy on both training and test sets, showing good performance and generalization.
Support Vector Machine: Similar performance to logistic regression, with good generalization.
k-Nearest Neighbors: Slightly lower accuracy, may need hyperparameter tuning.
Naive Bayes: Consistent performance, typically less affected by overfitting.
XGBoost, LightGBM, CatBoost: High accuracy on both training and test sets, demonstrating the effectiveness of gradient boosting methods.

4. Performance Analysis

Each model is trained on the training set and evaluate on the test set using the chosen metrics. Then the comparison of each algorithm is done on performance metrics

- Accuracy: The proportion of correct predictions (both true positives and true negatives) among the total number of cases.
- Precision: The proportion of true positive predictions among all positive predictions.
- Recall (Sensitivity): The proportion of true positives correctly identified by the model.
- F1-Score: The harmonic mean of precision and recall.
- ROC-AUC Score: The area under the Receiver Operating Characteristic curve, which plots the true positive rate against the false positive rate.

The following steps are followed:
(i) Load and Preprocess the data: Missing values are handled, categorical values are encoded and datas are split
(ii) Train the Models: Train each algorithm on the training data.
(iii) Evaluate the Models: Evaluate the models on the test data using various metrics.
(iv) Compare the Metrics: Compare accuracy, precision, recall, F1-score, and ROC-AUC

Table 3: Performance Metrics

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
K-Nearest Neighbors	0.82	0.83	0.84	0.83	0.87
Random Forest	0.89	0.87	0.93	0.90	0.94
Logistic Regression	0.85	0.83	0.89	0.86	0.90
Support Vector Machine (SVM)	0.87	0.85	0.91	0.88	0.92
Gradient Boosting Machine (GBM)	0.90	0.89	0.94	0.91	0.95

5. Conclusions

Heart disease is a dangerous condition that can lead to potentially catastrophic outcomes, including heart attacks, as well as potentially fatal complications. Using a heart disease dataset, this study aimed to evaluate the efficacy of ML techniques in cardiac disease prediction. The start of

cardiac disease can be accurately predicted by machine learning algorithms. To guarantee unbiased machine learning techniques, the data underwent pre-processing to address missing values, standardise the data, and address imbalances. Finally, a range of feature selection procedures were used to identify the most significant aspects that are highly advantageous for heart disease prediction, and six distinct machine learning algorithms were then applied to those features. Based on several sets of data, each algorithm produced a distinct score.

References

[1] World Health Statistics. Cardiovascular Diseases, Key Facts. 2021. Available online:room/factsheets/detail/cardiovascular-diseases-(cvds) (accessed on 10 December 2022).

[2] Choudhury, R.P.; Akbar, N. Beyond Diabetes: A Relationship between Cardiovascular Outcomes and Glycaemic Index. Cardiovasc.Res. 2021, 117, E97–E98.

[3] Magesh, G.; Swarnalatha, P. Optimal Feature Selection through a Cluster-Based DT Learning (CDTL) in Heart Disease Prediction.Evol. Intell. 2021, 14, 583–593.

[4] Rohit Chowdary, K.; Bhargav, P.; Nikhil, N.; Varun, K.; Jayanthi, D. Early Heart Disease Prediction Using Ensemble LearningTechniques. J. Phys. Conf. Ser. 2022, 2325, 012051.

[5] Liu, J.; Dong, X.; Zhao, H.; Tian, Y. Predictive Classifier for Cardiovascular Disease Based on Stacking Model Fusion. Processes2022, 10, 749.

[6] Devi, A.G. A Method of Cardiovascular Disease Prediction Using Machine Learning. Int. J. Eng. Res. Technol. 2021, 9, 243–246.

[7] Uddin, S.; Khan, A.; Hossain, M.E.; Moni, M.A. Comparing Different Supervised Machine Learning Algorithms for Disease Prediction. BMC Med. Inform. Decis. Mak. 2019, 19, 281.

[8] Patro, S.P.; Nayak, G.S.; Padhy, N. Heart Disease Prediction by Using Novel Optimization Algorithm: A Supervised Learning Prospective. Inform. Med. Unlocked 2021, 26, 100696.

[9] Pasha, S.J.; Mohamed, E.S. Novel Feature Reduction (NFR) Model with Machine Learning and Data Mining Algorithms for Effective Disease Risk Prediction. IEEE Access 2020, 8, 184087–184108.

[10] Deep learning for improving the effectiveness of routine prenatal screening for major congenital heart diseases. Nurmaini S, Partan RU, Bernolian N, et al. J Clin Med. 2022;11

[11] Azween Abdullah, S. Nithya, M. Mary Shanthi Rani, S. Vijayalakshmi and Balamurugan Balusamy, “Stacked LSTM and Kernel-PCA-based Ensemble Learning for Cardiac Arrhythmia Classification” International Journal of Advanced Computer Science and Applications(IJACSA), 14(9), 2023. http://dx.doi.org/10.14569/IJACSA.2023.0140905

[12] Tanwar, Sushama & Vijayalakshmi, S. & Sabharwal, Munish & Kaur, Manjit & Ali, Ahmad & Lee, Heung-No. (2022). Detection and Classification of Colorectal Polyp Using Deep Learning. BioMed Research International. 2022. 1-9. 10.1155/2022/2805607.

[13] Gokul Rajan V., S. Vijayalakshmi, Munish Sabharwal. (2021). Modified CNN based Feature Extraction for Sclera Recognition of Off . Annals of the Romanian Society for Cell Biology, 2579–2590.

[14] P. Durgadevi, S. Vijayalakshmi, Munish Sabharwal. (2021). Fetal Brain Abnormality Detection through PSO (Particle Swarm Optimization) and Volume Estimation. Annals of the Romanian Society for Cell Biology, 2700–2714