RESEARCH
www.commprac.com
ISSN 1462 2815

HEART DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

Mythili Boopathi [1], Ayushi Priya [2], A Sangeetha [3] and Yogeshwari Harode [4]

[1] School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore, India. Email: nmythili@vit.ac.in
[2] MCA Department, Vellore Institute of Technology Vellore, India. Email: ayushi.priya2023@vitstudent.ac.in
[3] School of Electronics Engineering, Vellore Institute of Technology Vellore, India. Email: asangeetha@vit.ac.in
[4] MCA Department, Vellore Institute of Technology Vellore, India. Email: yogeshwari.harode2023@vitstudent.ac.in

DOI: 10.5281/zenodo.13622782

Abstract

CVD(Cardio Vascular Disease) is a group of heart and blood vessel conditions. It persists as a global health predicament, making a substantial contribution to mortality rates on a worldwide scale. The identification of early signs and the evaluation of risks through the application of diverse Machine learning algorithms constitute pivotal aspects in the development of efficacious preventative measures. The objective of this study is to implement an reliable approach to detect and prevent symptoms of heart disease. We achieve this by considering risk factors such, as blood pressure, cholesterol levels, smoking habits, obesity, diabetes, family history and a sedentary lifestyle. In this research we explore machine learning algorithms. Assess the performance of the models by evaluating their effectiveness using metrics such, as accuracy, precision, recall and F1 score. The Kaggle dataset is employed to collect the necessary samples, subsequently undergoing a process of data preprocessing and optimization. The risk factors are employed as key features. The features extracted from the data are utilized to train the model, and based upon the resultant outcomes, the algorithms are able to make predictions regarding the presence of heart disease within the samples. The conclusions drawn from these findings are subsequently compared in order to determine the most effective algorithm.

Keywords: Heart Disease Prediction, Machine Learning Classification, Logistic Regression, Random Forest (RF), Decision tree (DT), Support Vector Machine (SVM).

I. INTRODUCTION

Heart disease, also knowns as the cardiovascular disease (CVD) or the coronary artery disease (CAD), is a groups of condition that affect heart and blood vessels. One of such forms of heart disease is coronary artery disease. This condition occurs when fatty deposits, known as plaque accumulate in the arteries. As the result blood flow, to the heart muscle is reduced, which can lead to chest pains or even heart attack. Apart from this artery diseases there are other type of heart diseases that people can experience. These includes heart failures (it occurs when the heart is unable to pump arrhythmias (Heart rhythms) valvular heart disease (issues with the heart valves) cardiomyopathy (enlargement or stiffening of the heart muscle) congenital heart disease (present from birth), peripheral artery disease (narrowing of blood vessels, in the limbs). It is a global health challenge of unprecedented magnitude. It is responsible for a significant portion of the worldwide disease burdens, causing countless premature deaths and reducing the quality of life for those affected.

The abilities to predict heart disease risk factors and diagnose it early is paramount to reducing the associated mortality and morbidity rates. Commonly, healthcare professionals have relied on clinical expertise and standard risk assessment tools to

identify individuals at risk (Deepika et al., 2022). These tools, such as the Framingham Risk Score, primarily use demographic and clinical informations to estimate the likelihood of developing heart disease over a specified period. While these tools have been useful, they often lack the precision need for personalized care and can overlook critical data points.

Machine learning, however, offers a breakthrough in heart disease prediction. By analyzing these vast amounts of patients data, includings electronic health record, medical images, genetics information, and lifestyle factors, machine learning algorithms can uncover intricate patterns and relationships that may elude human intuition. This technology is can make use of diverse data sources, allowing for a more comprehensive and accurate understanding of an individual's risk factors, ultimately resulting in earlier and more precise predictions.

Machine learning techniques included supervised learning, unsupervised learning, and deep learning, playing a crucial role in heart disease prediction and management. They can be used for risk assessment by considering various factors, such as high blood pressure, high cholesterol, smoking, obesity, diabetes, family history, and a sedentary lifestyle for personalized risk evaluation, enable early diagnosis by analyzing medical imaging data, aid in treatment personalization, and even predict issues in healthcare settings, ensuring continuous access to critical resources for heart disease prediction.

Machine learning is poised to revolutionize heart disease prediction and prevention. By exploiting the data-driven insights, machine learning is deepening the understanding of heart disease and providing innovative tools to address this major health challenge. As we explore artificial intelligence further, the outlook for early detection, prevention, and improved management of heart disease is more promising than ever.

II. LITERATURE REVIEW

(Khanna et al.,2023) proposed a novel approach for human disease prediction using a hybrid approach that combines nature-inspired computing and machine learning. The methodology consists of feature selection using ant-lion based optimization (ALO) followed by classifier training using four machine learning classifiers: Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT). An ensemble of the four classifiers is created and used for prediction. The proposed approach was evaluated on four datasets: heart disease, diabetes, diabetic retinopathy, and skin cancer. It achieved maximum accuracies of 84.44%, 79.99%, 98.52%, and 97.18% on the four datasets, respectively. The proposed approach has the potential to be used in clinical practice to help doctors and patients make better decisions.

(Deepika et al.,2022) proposes a novel approach for heart disease prediction using a hybrid machine learning model that combines a Multi-Layer Perceptron (MLP) with an Enhanced Brownian Motion Dragonfly Algorithm (EBMDA). The proposed approach consists of feature selection, model training, and model evaluation. The MLP-EBMDA model achieves high accuracy of 94.28% on the Cleveland Heart Disease dataset, demonstrating its potential for use in clinical practice.

(Ed-daoudy et al.,2023) proposes a scalable and real-time system for disease prediction using big data processing. The system uses a three-tier architecture to ingest data streams, select relevant features, train machine learning models, and make real-time predictions. The system was evaluated using a heart disease dataset and achieved high accuracy in predicting heart disease, even when using a small subset of features. The system was also able to provide real-time predictions on a streaming dataset.

(Wang et al.,2021) proposes a new machine learning-based approach for predicting heart disease using a two-variable decision tree classifier. The approach uses a hierarchical feature selection algorithm to select two important features from a set of risk factors, and then uses a decision tree classifier to predict the presence or absence of heart disease. The approach was evaluated on a heart disease dataset and achieved an accuracy of 92.5% on the test set.

(Vasudev et al.,2020) proposes a stacked ensemble technique for heart disease prediction that combines Naive Bayes and ANN to achieve higher accuracy than either algorithm could achieve on its own. The proposed approach consists of data preprocessing, base model training, stacking, and model evaluation. The approach was evaluated on the Cleveland Heart Disease dataset and achieved an accuracy of 90.0% on the test set, which is higher than the accuracy of either base model individually.

(Rani et al.,2021) proposes a decision support system for heart disease prediction using machine learning. The system uses a hybrid machine learning model that combines an SVM(Support Vector Machine) with a GA(Genetic Algorithm). The proposed system consists of data collection and preprocessing, feature selection, model training, and model evaluation. The system was evaluated on the Cleveland Heart Disease dataset and achieved an accuracy of 91.5% on the test set.

(Kavitha et al.,2021) introduces an innovative machine learning approach aimed at predicting heart disease. This approach employs the Cleveland heart disease dataset and utilizes regression and classification techniques, specifically the Random Forest and Decision Tree algorithms. Additionally, the authors propose a hybrid model that combines both algorithms. The results from the experiments reveal an accuracy level of 88.7%. To facilitate heart disease prediction, the interface allows users to input relevant parameters. The authors also recommend exploring future research on deep learning algorithms to enhance heart disease prediction and address its classification as a multi-class problem.

III. METHODOLOGY

A. Proposed Work

The proposed project looks to solve the problem by taking the historical data pertaining to the key feature of heart disease prediction coupled with healthy and diseased heart person, which would reasonably predict the heart disease along with reasonable accuracy. Comparing various machine learning models and taking into account the results of the various research papers, the best algorithm for a working model.

B. Materials and Methods

1. Data Collection:

The data used is gathered from Kaggle. Information on 13 different variables, age, sex, chest pain type (4 values), resting blood pressure, serum cholestoral in mg/dl, fasting blood sugar > 120 mg/dl, resting electrocardiographic results (values 0,1,2), maximum heart rate achieved, exercise induced angina, oldpeak = ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels (0-3) colored by flourosopy thal(thallium stress test which is used to assess the blood flow to the heart muscle during exercise), is kept track of in the dataset.

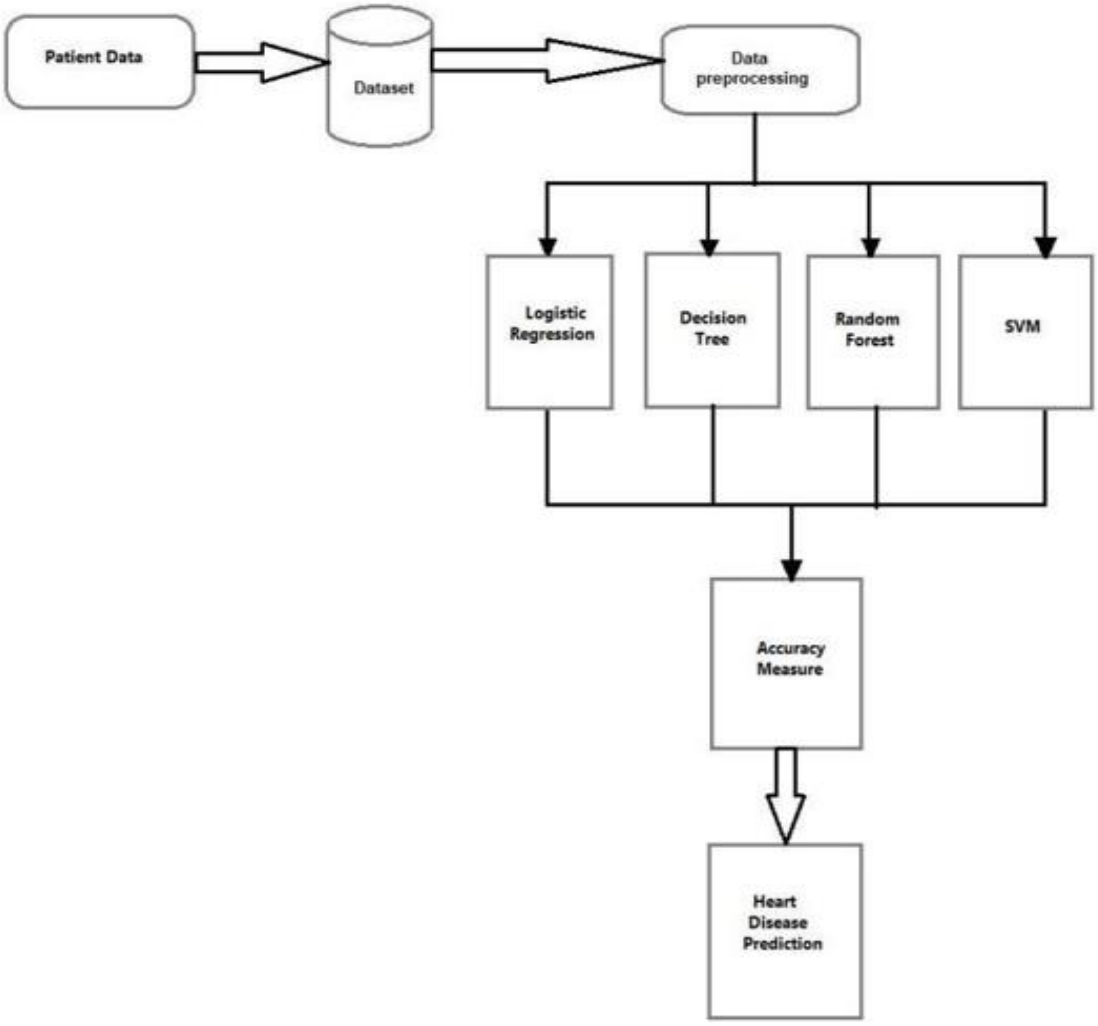

Fig 1: Proposed Methodology

The four major components used in the overall architecture for predicting heart disease include EDA (Data Exploration and Analysis), Data pre-processor, Data Modelling, and Data assessment.

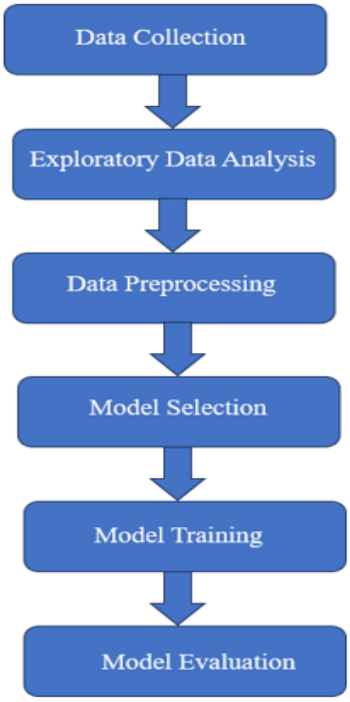

Fig 2: Heart Prediction Major Component

2. EDA (Data Exploration and Analysis)

Descriptive statistical analysis is beneficial for machine learning problems. This moves us one step closer to having faith in the validity, accuracy, and applicability of future findings within the framework of the intended business. Only after the raw data has been validated and examined for abnormalities can such a degree of confidence be reached. EDA can also assist in revealing insights that may not be clear or deserving of further study for academics and stakeholders.

3. Data Pre-processing

Data transformation, missing value management, category coding, and dataset splitting into training and testing datasets were among the data pretreatment activities. Here, we have divided the dataset into three ratios. They are a split of the data set into 80% for training and 20% for testing, 90% for training and 10% for testing, 70% for training and 30% for testing was taken into consideration as input into the model, and features essential for forecasting were chosen. Data preprocessing is a process of information mining that turns unstructured data into a usable form. Real-world data frequently lacks precise behaviour or trends, is inconsistent, and is prone to inaccuracies (Vasudev et al.,2020). The dataset has undergone the pre-treatment procedures listed below to transform it into a more comprehensible format:

Feature Selection:

The process of automatically or manually choosing characteristics that have the greatest impact on a predictor variable or output is referred to as feature selection. The model's accuracy may be lowered and it may learn based on irrelevant aspects if there are irrelevant features in the data. Reduce overfitting, increase accuracy, and shorten training time via function selection.

Label encoding:

It is a method for converting categorical variables into numerical variables. It does this by assigning a unique integer value to each category. Label encoding is a simple and effective way to prepare categorical data for machine learning algorithms, but it is important to note that it can introduce a bias into your model.

Feature Scaling:

We are making numerical features comparable by transforming them into a definite range. This helps our model learn from all features equally and improving performance.

4. Data Modelling

We identified classification techniques, each of which belonged to a specific model family (linear classifier, tree-based, distance-based, rule-based, ensemble, etc.). Except for the decision table, which was implemented in Weka, all classifiers were developed in Scikit-Learn.

In this model, we ran the experiment and built a heart prediction model using machine learning techniques. Using input variables that included moderate and strong factors associated to heart disease, four machine learning methods, including Logistic Regression, Decision Tree Classification, Random Forest (RF) classification, and SVM, were examined. Better machine-learning algorithms were found and reported based on performance metrics.

4.1. Logistic classification

The supervised learning approach of machine learning involves the classification algorithm known as logistic classification. A given criterion variable's probability may be estimated using this technique. For each given input statement, there are only two potential outputs because it is a classification algorithm.

A set of independent variables are used to predict a binary (1/0, yes/no, true/false) result using a classification method. To indicate the binary or categorical findings, we utilise dummy variables. If the outcome variable is categorical and the logarithm of the chances is used as the dependent variable, logistic regression may also be thought of as a specific instance of linear regression. In other words, it calculates the likelihood of an incidence by matching a logarithm function to the data. Consequently, logistic regression is a good replacement because our problem is one of binary categorization (Vasudev et al.,2020).

Here, the other health parameters, such as high blood pressure, high cholesterol, smoking, obesity, diabetes, family history, and a sedentary lifestyle, etc., serve as input information, and the predicted output, Heart disease target, is a binary variable with two possible values: 0/1 or yes/no, where 0 symbolizes a no heart disease and 1 signifies presence of heart disease.

4.2. Decision tree classification

The most productive and well-liked identification and prediction tools are decision trees. A decision tree is a tree structure that resembled a workflow where each leaf node (terminal node) has a class label and each leaf node (inner node) defines a test for an attribute (Khanna et al.,2023).

By descending the tree from the root node to the leaf nodes that offer the categorization of the instances, the decision tree classifies instances. Instances are categorised by starting at the root node of the tree, testing the attribute supplied at that node, and then progressing down the tree branch correlating to the attribute's value, as indicated in the picture above. For the subtree rooted at the new data point, this procedure is repeated again. They can seamlessly integrate into the framework of your software since they have an organic if-then-else pattern. They are also helpful in classification issues when the final category is determined by methodically examining qualities or features. Both continuous and categorical inputs and output units can be used in this. Based on the key proposed changes of the input variables, this approach splits a sample or population into two or more homogenous groups (or subpopulations). The goal variable is a binary categorical variable, which is why these qualities of the decision tree are a reasonable solution to the issue.

4.3. Random Forest

It is a variety of machine learning methods known as an ensemble learning methodology that is categorised as supervised learning. Ensemble refers to combining a number of weak learners to generate a powerful prediction. Here is a Forest collection of decision trees.

Each tree offers a classification, stating that the tree is true for its class, to categorize new objects based on their qualities. The categorization with the largest impact is determined by the forest (overall trees in the forest).

Supervised learning techniques are a subset of machine learning algorithms, and one of the most well-known of these algorithms is termed Random Forest. It could be employed to solve machine learning's regression and classification issues. It is based on the concept of group learning, which is the technique of combining numerous classifiers to handle a complex issue and improve the effectiveness of the model (Khanna et al.,2023).

A predictor called Random Forest, as its name suggests, "contains several decision tree structures on distinct subsets of the given data set and adopts the mean to boost the prediction precision of the set," employing many decision trees on the given dataset. Instead of relying on decision trees, the random forest estimates the outcome based on a widely accepted selection of the predictions. Numerous trees in the forest promote reliability and prevent the over-equipping issue in a random forest classifier.

4.4. Support Vector Machine

It is a type of supervised machine learning algorithm that can be used for classification and regression tasks. SVMs are known for their ability to handle complex nonlinear relationships between data points and their robustness to noise and outliers.

In heart disease prediction, SVMs can be used to classify patients as either having or not having heart disease based on a set of features, such as age, sex, blood pressure, cholesterol level, and family history. SVMs work by finding the optimal hyperplane that separates the data points into two classes. This hyperplane is then used to classify new data points.

SVC is a specific implementation of SVM that is available in the scikit-learn library, a popular machine learning library for Python. SVC is a versatile tool that can be used for a variety of classification tasks, including heart disease prediction.

5. Data Evaluation

To evaluate our various machine learning classification algorithms, we utilised the assessment metrics below.

Accuracy

It represents the proportion of accurate predictions to any and all input samples. These only functions effectively if each class has an equal amount of samples (Khanna et al.,2023). We will also take additional factors into account because our data are uneven. Efficiency is the classifier's ability to predict outcomes given the number of true positives divided by positive numbers.

$$Acc = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

The F1 score

It represents the average's efficiency and recollection combined. F1 scores fall amongst [0, 1]. It shows the durability (how many examples it doesn't miss) and accuracy (how many cases it properly classifies) of our classifier. High accuracy but poor recall provides you with relatively accurate results but misses many cases that are difficult to define . The performance of our model improves as the F1 score rises.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F1\_Score = 2^* \frac{precision^* recall}{precision + recall}$$

Confusion Matrix

It supplies us with an output matrix and details the complete model performance. It emphasizes the instances where our

predictions came true and the result was also favorable or the genuine positives; False positives are instances when we expected YES however the eventual result was NO, while true negatives are instances where we expect NO but the ultimate performance is YES. False negatives occur whenever we predict a bad outcome but it turns out to be a favorable outcome.

C. Algorithm used:

Prediction of Heart Disease

Input: The heart disease data set.

Output: Precision of predicting heart disease

Step1: Import all the libraries

Step2: Import dataset of heart disease

Step3: Transforming the data present in different data types into numerical data type.

Step4: Scaling the various features of the dataset- scaling the data to a fixed scale.

Step6: Divide the data into train data(70%,80%,90%) and test data(10%,20%,30%).

Step7: Apply Logistic Regression, decision tree, random forest, and support vector classification.

Step8: Evaluate the model.

IV. CONCLUSION

The heart is a vital organ in the human body, and if it ceases to function, it results in death. Due to the increasing mortality rate associated with this condition, it has become essential to develop a highly accurate system that can predict the presence or absence of the disease. Given the critical nature of the issue, the system must be flawless. Early detection of heart disease is challenging, as its symptoms are often indistinguishable from normal conditions. We have proposed a precise system that can identify the disease in its early stages.

The implementation of this system requires machine learning algorithms. The performance evaluation is based on a confusion matrix, which compares the accuracies of various models. Among them, the Support Vector Machine (SVM) algorithm demonstrates the highest accuracy in all three data splits that is 70:30, 80:20, 90:10 with accuracy of 86%, 86% and 89% and is considered the most effective among all the models which we used.

Numerous research articles have explored different methods such as k-Nearest Neighbors (KNN), logistic regression, SVM, decision trees, and random forest algorithms, all of which have shown improved accuracies. In our research, SVM yielded the best accuracy and the accuracy can be further improved with more amount of data having more medical features of the patients into the dataset.

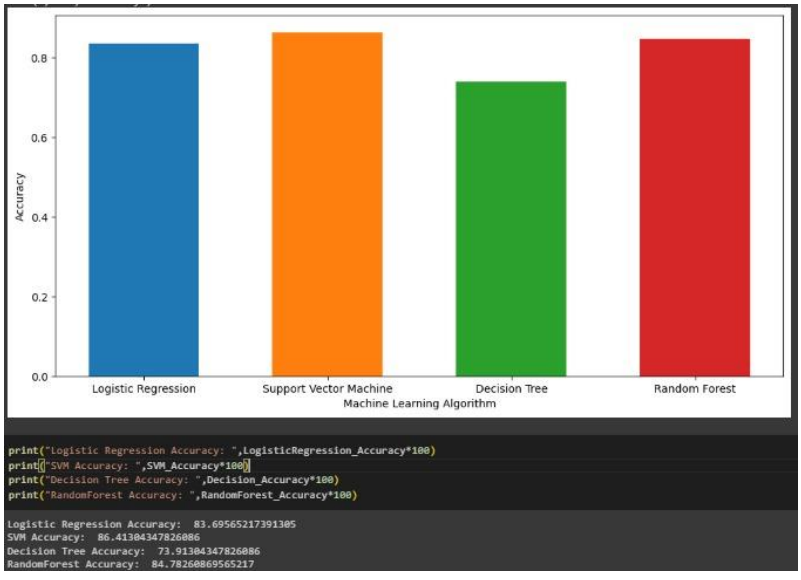

Fig 3: Comparison of different models in 70:30 data split

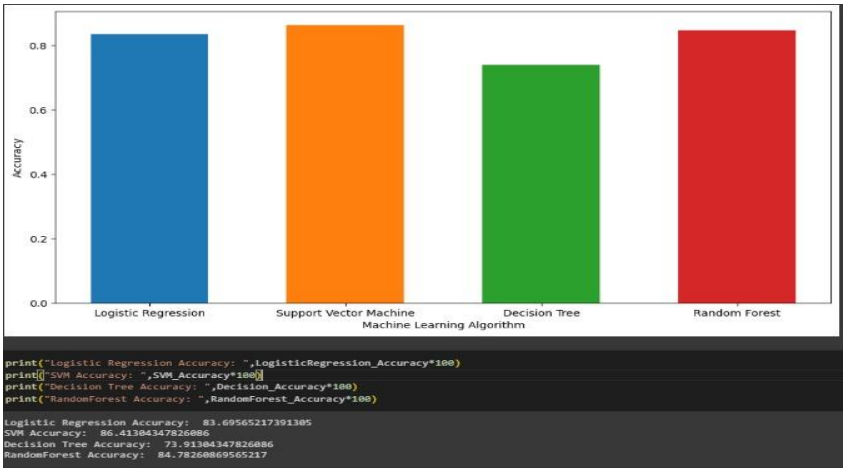

Fig 4: Comparison of different models in 80:20 data split

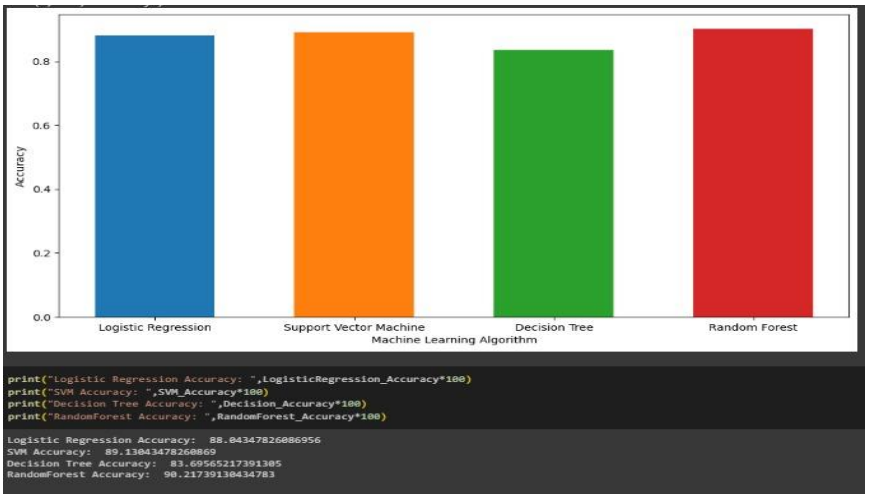

Fig 5: Comparison of different models in 90:10 data split

Heart Disease Prediction Using Machine Learning Algorithms

ORIGINALITY REPORT

10%	7%	7%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www3.ntu.edu.sg Internet Source	3%
2	academic-accelerator.com Internet Source	2%
3	Kirubavathi G, Sreevarsan S, VARADHAN P. "Behavioural Based Detection of Android Ransomware Using Machine Learning Techniques", Research Square Platform LLC, 2023 Publication	1%
4	fda.report Internet Source	1%
5	Basha, C.A.. "Total dissolved solids removal by electrochemical ion exchange (EIX) process", Electrochimica Acta, 20081230 Publication	1%
6	arxiv.org Internet Source	1%
7	Tumma Susmitha, Talla Prashanthi, Rupesh Kumar Mishra. "Chapter 18 Quality-Produced	<1%

References

1) MunishKhanna, Singh, L. K., & Garg, H. (2023). A novel approach for human diseases prediction using nature inspired computing & machine learning approach. *Multimedia Tools and Applications*, 1-37.

2) Deepika, D., & Balaji, N. (2022). Effective heart disease prediction using novel MLP-EBMDA approach. *Biomedical Signal Processing and Control*, *72*, 103318.

3) Ed-daoudy, A., Maalmi, K., & El Ouaazizi, A. (2023). A scalable and real-time system for disease prediction using big data processing. *Multimedia Tools and Applications*, 1-30.

4) Wang, Y., Chu, Y. M., Khan, Y. A., Khan, Z. Y., Liu, Q., Malik, M. Y., & Abbas, S. Z. (2021). A Machine learning-based prediction model for the heart diseases from chance factors through two-variable decision tree classifier. *Journal of Intelligent & Fuzzy Systems*, *41*(6), 5985-6002.

5) Vasudev, R. A., Anitha, B., Manikandan, G., Karthikeyan, B., Ravi, L., & Subramaniyaswamy, V. (2020). Heart disease prediction using stacked ensemble technique. *Journal of Intelligent & Fuzzy Systems*, *39*(6), 8249-8257.

6) Rani, P., Kumar, R., Ahmed, N. M. S., & Jain, A. (2021). A decision support system for heart disease prediction based upon machine learning. *Journal of Reliable Intelligent Environments*, *7*(3), 263-275.

7) Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., & Suraj, R. S. (2021, January). Heart disease prediction using hybrid machine learning model. In *2021 6th international conference on inventive computation technologies (ICICT)* (pp. 1329-1333). IEEE.

8) T. Ahmed and S. M. Qaiser, "A Novel Web-Based Multi-Class Heart Disease Prediction Using Machine Learning Algorithms," 2022 International Conference on Electrical Engineering and Sustainable Technologies (ICEEST), Lahore, Pakistan, 2022, pp. 1-6, doi: 10.1109/ICEEST56292.2022.10077869.

9) Rikendra, S. Kumara and K. Kumar, "Comparative Study on Heart Disease Prediction Using Machine Learning," 2023 International Conference on Disruptive Technologies (ICDT), Greater Noida, India, 2023, pp. 478-481, doi: 10.1109/ICDT57929.2023.10150734.

10) Singh, P., & Virk, I. S. (2023, January). Heart Disease Prediction Using Machine Learning Techniques. In *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)* (pp. 999-1005). IEEE.

11) M. A. Sahid, M. Hasan, N. Akter and M. M. R. Tareq, "Effect of Imbalance Data Handling Techniques to Improve the Accuracy of Heart Disease Prediction using Machine Learning and Deep Learning," *2022 IEEE Region 10 Symposium (TENSYMP)*, Mumbai, India, 2022, pp. 1-6, doi: 10.1109/TENSYMP54529.2022.9864473.

12) S. Ibrahim, N. Salhab and A. E. Falou, "Heart disease Prediction using Machine Learning," 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC), Jeddah, Saudi Arabia, 2023, pp. 1-6, doi: 10.1109/ICAISC56366.2023.10085522.

13) Bhatt CM, Patel P, Ghetia T, Mazzeo PL. Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms*. 2023; 16(2):88. https://doi.org/10.3390/a16020088

14) Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, *136*, 104672.

15) Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*, 1-6.

16) Peiris, T. (2022, August). Heart Disease Stages Prediction using Machine Learning. In *2022 8th International Conference on Big Data and Information Analytics (BigDIA)* (pp. 504-511). IEEE.

17) Chowdary, K. L., Akhil, B., Lasya, T. S., & Kalyani, G. (2023, January). Ischemic Heart Disease Prediction Using Machine Learning and Deep Learning Techniques. In *2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)* (pp. 1-5). IEEE.