SMART HEART DISEASE PREDICTION USING MACHINE LEARNING

Ms. V. K. Monisha
Assistant Professor, Department of Computer Science
Hindustan institute of technology and science, Chennai, Tamil Nadu, India

Ms. D. Kamalin
Assistant Professor, Department of Computer Science
Hindustan institute of technology and science, Chennai, Tamil Nadu, India

Ms. T.Vani
Assistant Professor, Department of Computer Science
Hindustan institute of technology and science, Chennai, Tamil Nadu, India
DOI: doi.org/10.34293/seamjr-v1i3-012

Abstract

Heart Disease (HD) is among the maximum prevalent diseases today, and to a healthcare- providing society, it is extremely important to collaborate with them to maintain their patients' health and save their lives. In this plan, various classifiers were tested based on performance comparison to classify the heart disease dataset to classify it accurately and/or to Predict Heart Disease cases with fewer features. The machine learning activity can have an important part to play in the healthcare industry. The healthcare industry collects a large amount of medical information, and it is impossible to find the unseen information to make effective decisions. Prediction system of Heart disease through more input features. The method incorporates medical words such as blood pressure, cholesterol, and sex. Thirteen factors, including family history, smoking, poor diet, high blood pressure, high blood cholesterol, obesity, physical inactivity, and hypertension, might predict a patient's risk of heart disease prediction Based on our outcome's accuracy of Decision Tree Support Vector machine, and random forest respectively. Our evaluation demonstrates that among these three classification models support vector machines predict heart disease with the best accuracy. Early results demonstrate high prediction accuracy, which means machine learning may render heart disease prediction more accessible and effective.

***Keywords**: Heart Disease Prediction, Machine Learning, Predictive Analysis, Random Forest (RF), Support Vector Machine (SVM), Decision Tree*

Introduction

Machine learning as a buzzword of the last few years, perhaps the cause lies in the immense amount of data generation by application, the rise of totaling power Over the last few years, and the emergence of improved algorithms. Every industry aims to gain from machine learning, which is used for everything from automating repetitive tasks to generating insightful information. It may already be a feature of a product you use. For example, a smart home gadget like Google Home or a wearable fitness and health tracker like Fitbit. There are numerous additional instances of ML. Heart disease causes approximately 31% of global deaths each year, fueled by conditions such as sedentary lifestyles, diets, stress, and genetic susceptibilities. Timely diagnosis is important to avoid complications and enhance quality of life. Limitations of Traditional Diagnostic Techniques:

Conventional diagnostics (e.g., ECGs, stress tests) are complex and need trained staff and specialized equipment, making them less accessible in resource-constrained environments and causing delays in treatment and diagnosis. Machine learning has the potential to benefit healthcare by comparing large amounts of patient information to find patterns that assist in early diagnosis and prevention, with the potential to relieve healthcare system pressures. Random Forest, an ensemble technique, is good at dealing with large, complex data sets, giving insights into feature significance (e.g., cholesterol, blood pressure), and minimizing overfitting to enable model generalization to other patient populations.

Support Vector Machine (SVM) is good at binary classification, especially in high-dimensional data sets, thus enabling the use of heart disease prediction in discrimination between "at risk" and "not at risk" individuals. The integration of Random Forest's feature importance evaluation and SVM's acute classification borders makes a stable prediction framework, allowing clinicians to make educated, data-based choices and assisting proactive medicine with the early identification of risks.

Literature Review

In this paper [1] Coronary Heart Disease (CHD) is still among the top morbidity and mortality causes in the world, and as such, it is necessary to develop early detection and prediction for enhanced patient outcomes. This experimental study analyzes the use of different machine learning (ML) algorithms to predict CHD risk. The research employs a dataset with vital health indicators, such as age, cholesterol levels, blood pressure, smoking, and family heart disease history, to train and test various ML models like logistic regression, decision trees, random forests, SVM, and neural networks. The experimental findings show that ensemble approaches such as random forests and neural networks deliver the best predictive performance, whereas more simple models such as logistic regression yield a trade-off between interpretability and accuracy.

In this paper [2] A survey on heart disease prediction methods by machine learning can investigate several different approaches and models that have been utilized for heart disease prediction possibility based on several different risk factors, including age, gender, cholesterol level, blood pressure, smoking status, exercise, and history of heart disease. Here is a summary of some of the major machine learning methods that are popularly applied to heart disease prediction: Logistic regression is employed for predicting the likelihood of a categorical dependent variable. In heart disease prediction, it is utilized for determining based on a person's features whether they are at risk for heart disease.

This paper [3] suggests a predictive analysis framework for heart disease that uses machine learning techniques to identify key variables and predict a patient's likelihood of developing heart disease. Several machine learning algorithms, including logistic regression, are used by the framework. Random Forest, Support Vector Machine (SVM) and Neural Networks, to investigate various health parameters such as blood pressure, age, cholesterol, and ancestral histories of heart disease.

In [4] this article Coronary Heart Disease (CHD) is still a major global health concern, and early detection is essential to reducing morbidity and death. Using machine learning algorithms on the Cleveland dataset, which contains patient data on a variety of cardiovascular risk factors like age, sex, blood pressure, cholesterol levels, electrocardiographic findings, and exercise- induced angina, the current study aims to predict coronary heart disease (CHD) early. Data cleaning, normalization, and feature selection are examples of preprocessing tasks that are carried out to enhance the models' performance.

In this article [5] The ability to predict and diagnose heart disease early is crucial to healthcare since it remains one of the leading causes of death worldwide. The use of machine learning techniques to predict heart disease based on various clinical parameters, such as age, blood pressure, cholesterol, electrocardiogram results, and family history, is investigated in this study.

According to this paper [6], early diagnosis can significantly improve treatment outcomes and lower death rates, making heart disease prediction a significant healthcare issue. This study will investigate the potential of machine learning techniques in the prediction of heart disease based on biomedical data, including clinical characteristics like age, sex, blood pressure, cholesterol level, ECG data, angina brought on by exercise, and other comparable health indicators

The Cleveland Heart Disease dataset offers a comprehensive set of patient data, while other publicly available sources provided the data used in this analysis. To improve model performance, crucial preprocessing methods like data cleaning, feature selection, and normalization are used. The effectiveness of machine learning algorithms in predicting heart disease is demonstrated by this study, which has significant implications for early intervention and clinical judgment. In order to facilitate prompt diagnosis and preventive care, the findings demand that machine learning tools be incorporated into health systems.

In this article [7] Early detection is crucial in lessening the impact of heart disease, which is one of the leading causes of death worldwide. Through a hybrid machine learning model that leverages the advantages of multiple algorithms to increase prediction accuracy and reliability, this study provides a useful approach for heart disease prediction. It seeks to improve predictive performance by balancing their distinct advantages and disadvantages. The Cleveland Heart Disease dataset, which includes several health parameters like age, sex, blood pressure, cholesterol, exercise-induced angina, and other significant factors, is used in this study. Data preprocessing techniques such as feature selection, missing value handling, and normalization are used to prepare the data for model training.

The results show that the hybrid machine learning model outperforms single models in terms of accuracy and generalization. It is a very promising method for the early detection and prevention of heart disease because the strength of using multiple algorithms increases the overall predictive strength. This research highlights the applicability of using hybrid machine learning models to enable better healthcare through more precise and dependable predictions for timely decision-making and intervention strategies for clinicians.

In this paper [8] This work suggests a Heart Disease Identification Method based on machine learning classification methods, specifically created for embedding within healthcare systems. The work intends to create an automated, efficient, and precise model for the diagnosis of heart disease based on biomedical information such as age, gender, cholesterol level, blood pressure, family history, electrocardiogram readings, and exercise-induced angina, among others. Data preprocessing methods like feature scaling, data cleansing, and filling missing values are included in the process to provide high-quality inputs to the models.

This article [9] One of the leading causes of death worldwide is still heart disease, and reducing risks and improving patient outcomes depend on early detection. This paper offers a comprehensive review of several machine learning techniques used in heart disease prediction, assessing their benefits, drawbacks, and applications. The use of several algorithms, such as ensemble learning techniques, K- Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest, Decision Trees, Logistic Regression, and Artificial Neural Networks (ANN), for the identification of heart disease risk factors is covered in the article. The preprocessing steps involved in the survey include data cleaning, feature scaling, normalization, and missing value handling, which are key factors influencing the performance of the predictive models.

The findings of several studies indicate that algorithms like Random Forest and SVM tend to perform better than other algorithms in prediction accuracy and reliability. This survey concludes by highlighting the promise of machine learning in enhancing the early diagnosis of heart disease, allowing for timely intervention and improved management of patient health.

In this paper [10] Forecasting heart rates is an important element in determining cardiovascular well-being and controlling conditions like arrhythmia, hypertension, and heart disease. The research here seeks to present a predictive heart rate analysis based on many machine learning methods using biomedical data from age, weight, activity level, blood pressure, and past heart rate records. The objective is to build a model that will predict a person's heart rate during various physiological states so that healthcare practitioners are better able to monitor and take care of a patient's cardiovascular health. The data for analysis comprises time-series records of heart rate values obtained from wearable sensors and clinical trials, with a variety of inputs to reflect real-world conditions. The data for analysis comprises time-series records of heart rate values obtained from wearable sensors and clinical trials, with a variety of inputs to reflect real-world conditions. Preprocessing steps involving missing value handling, normalization, and feature extraction are carried out to enhance the performance of the models. The models are then tested based on critical performance indicators, including mean squared error (MSE), accuracy, and R-squared to see how accurately they can predict and withstand real-world situations. The results show that ensemble models such as Random Forest and neural networks provide better performance in heart rate prediction, especially in terms of capturing non-linear relationships among features. This predictive model offers a valuable tool for medical applications, allowing for constant monitoring and early detection of abnormal heart rates and hence ensuring personalized health care. The research also points to future developments in the integration of real-time monitoring systems like wearables with machine learning models for better and timelier heart rate prediction.

Methodology

To build an automatic predictive model based on Decision tree, Random Forest, and SVM-based predictors for the risk of heart disease in predicting the values with historical patient data. Determine and compare the performance of the Random Forest and SVM models to be able to identify which one is best suited for the prediction of heart disease in terms of factors like accuracy, precision, recall, and F1 score for clinical purposes. To analyze and rank the clinical and demographic features, that may result in heart disease, such as cholesterol, high blood pressure, age, lifestyle, etc.,

Support Vector Machine

To achieve classification, SVM optimizes the margin between two classes by subtracting the hyperplane. The support vectors are the vectors (cases) that define the hyperplane. Support vector networks, or SVMs, are supervised learning models in machine learning that scan data for regression and classification using related learning methods.

SVM Type 1 Classification

The error function is minimized to prepare for such an SVM:

$$\frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i$$

Subject to the constraints:

$$y_i \left(w^T \phi(x_i) + b \right) \geq 1 - \xi_i \text{ and } \xi_i \geq 0,\ i = 1, \ldots, N$$

This represents parameters for processing no separable data (inputs), where C is the continuous capacity, w is the vector of coefficients, and b is a constant. N training cases are designated by index i. represents the independent variable, and xi represents the class label. Data is mapped from the input (independent) to the article space using the kernel. It should be mentioned that the punishment for the error increases with the C. Therefore, to avoid overfitting, C should be carefully chosen.

Random Forest

A new deep learning algorithm that is categorized as a controlled learning technique is called Random Forest. It can be applied to DL issues involving both classification and regression. It is based on the concept of ensemble learning, which is the process of combining numerous classifiers to improve the model's performance and solve a challenging issue.

Random Forest Prediction:

Classification: The final output of the Random Forest model is determined by majority voting (or mode) of all the decision trees

$$\hat{y} = \text{mode}\left(f_1(x), f_2(x), \ldots, f_K(x)\right)$$

Where

$$f_1(x), f_2(x), \ldots, f_K(x)$$

are the predictions from each tree.

Random Forest's Formula in General Terms

$$\hat{y} = \frac{1}{K}\sum_{k=1}^{K} f_k(x) \quad \text{(for regression)}$$

Decision Tree

Decision tree is a supervised learning algorithm used for both classification and regression tasks. It works by dividing data into subsets according to feature values, with each division leading to branches that are decisions. We employ a method to construct a decision tree that splits the data in a manner that subsets are as "pure" as possible. For classification, the most frequently utilized criterion is Gini Impurity or Entropy (Information Gain). For regression, we may utilize the Mean Squared Error (MSE). 1Gini Impurity (for classification)

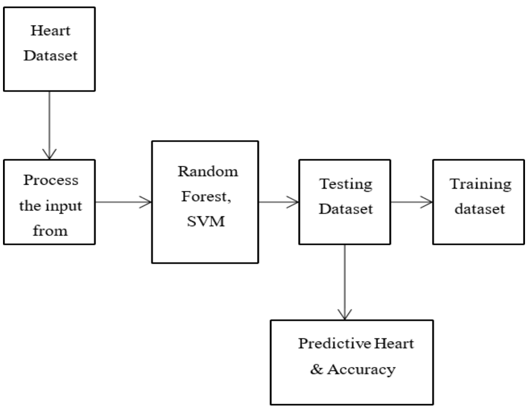
The Gini Impurity of a node t is calculated as:

$$Gini(t) = 1 - \sum_{i=1}^{c} p_i^2$$

Where:

- c is the number of classes.
- p_i is the probability of the class iii at node t

Architecture Diagram



Experimental Results

In order to predict Heart Disease from the given data set instances, this study uses classification methods such as the SVM algorithm and the Decision Tree Random Forest algorithm. The performance of the proposed methods is assessed using the Python Flask tool on the Heart Disease dataset. The main organ in the human body is the heart.

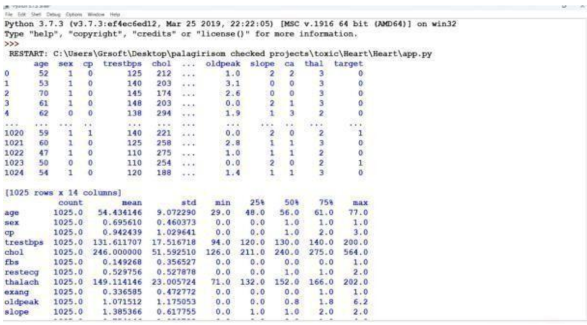
Dataset

The open-heart disease database is used. There are 1026 records in the Heart Disease database and 1026 records in the Stat log Heart Disease database [12]. The set includes the following five types of attributes: input, predictive, key, and other types.

Dataset Description:

- Age: The patient's age (int).
- sex: The patient's gender; 0 for female (int), 1 for male.
- cp: kind of chest pain (int); atypical angina, nonanginal pain, atypical angina, and asymptomatic are the four possible outcomes.
- The resting blood pressure (mm Hg) (int) is known as trestbps.
- chol: Level of serum cholesterol (mg/dl)
- Fasting blood sugar (FBS) > 120 mg/dl; true = 1, false = 0.
- resting ECG findings (int); 0: Normal, 1: Abnormal ST-T wave, and 2: Indicates suspected or confirmed left ventricular hypertrophy.
- Thalach: The highest heart rate attained (int).
- exang: Exercise-induced angina; 0 (int) if not, 1 otherwise.
- Oldpeak: Depression brought on by physical activity as opposed to rest (float).
- The slope of the peak workout ST segment (int) can be classified as either upsloping, flat, or downsloping.
- ca: The number of significant vessels that fluoroscopy has colored (float).
- Thalassemia; 3: Normal; 6: Defect Fixed; 7: Defect Reversible (float).
- num: Diagnosis of heart illness (int); 0: No disease, 1-4: Increasingly severe disease

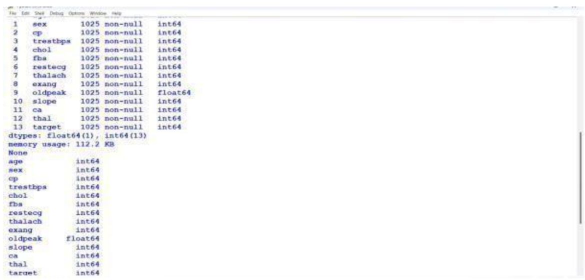
Result Analysis Main Page



```
Python 3.7.3 (v3.7.3:ef4ec6ed12, Mar 25 2019, 22:22:05) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
 RESTART: C:\Users\Grsoft\Desktop\palagirisom checked projects\toxic\Heart\Heart\app.py
      age  sex  cp  trestbps  chol  ...  oldpeak  slope  ca  thal  target
0      52    1   0       125   212  ...      1.0      2   2     3       0
1      53    1   0       140   203  ...      3.1      0   0     3       0
2      70    1   0       145   174  ...      2.6      0   0     3       0
3      61    1   0       148   203  ...      0.0      2   1     3       0
4      62    0   0       138   294  ...      1.9      1   3     2       0
...   ...  ...  ..       ...   ...  ...      ...    ...  ..   ...     ...
1020   59    1   1       140   221  ...      0.0      2   0     2       1
1021   60    1   0       125   258  ...      2.8      1   1     3       0
1022   47    1   0       110   275  ...      1.0      1   1     2       0
1023   50    0   0       110   254  ...      0.0      2   0     2       1
1024   54    1   0       120   188  ...      1.4      1   1     3       0

[1025 rows x 14 columns]
           count        mean        std    min    25%    50%    75%    max
age       1025.0   54.434146   9.072290   29.0   48.0   56.0   61.0   77.0
sex       1025.0    0.695610   0.460373    0.0    0.0    1.0    1.0    1.0
cp        1025.0    0.942439   1.029641    0.0    0.0    1.0    2.0    3.0
trestbps  1025.0  131.611707  17.516718   94.0  120.0  130.0  140.0  200.0
chol      1025.0  246.000000  51.592510  126.0  211.0  240.0  275.0  564.0
fbs       1025.0    0.149268   0.356527    0.0    0.0    0.0    0.0    1.0
restecg   1025.0    0.529756   0.527878    0.0    0.0    1.0    1.0    2.0
thalach   1025.0  149.114146  23.005724   71.0  132.0  152.0  166.0  202.0
exang     1025.0    0.336585   0.472772    0.0    0.0    0.0    1.0    1.0
oldpeak   1025.0    1.071512   1.175053    0.0    0.0    0.8    1.8    6.2
slope     1025.0    1.385366   0.617755    0.0    1.0    1.0    2.0    2.0
```

Fig (a): Dataset Pre-processing



```
 1   sex       1025 non-null   int64
 2   cp        1025 non-null   int64
 3   trestbps  1025 non-null   int64
 4   chol      1025 non-null   int64
 5   fbs       1025 non-null   int64
 6   restecg   1025 non-null   int64
 7   thalach   1025 non-null   int64
 8   exang     1025 non-null   int64
 9   oldpeak   1025 non-null   float64
 10  slope     1025 non-null   int64
 11  ca        1025 non-null   int64
 12  thal      1025 non-null   int64
 13  target    1025 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
None
age            int64
sex            int64
cp             int64
trestbps       int64
chol           int64
fbs            int64
restecg        int64
thalach        int64
exang          int64
oldpeak      float64
slope          int64
ca             int64
thal           int64
target         int64
```

Fig (b): Dataset Datatype details

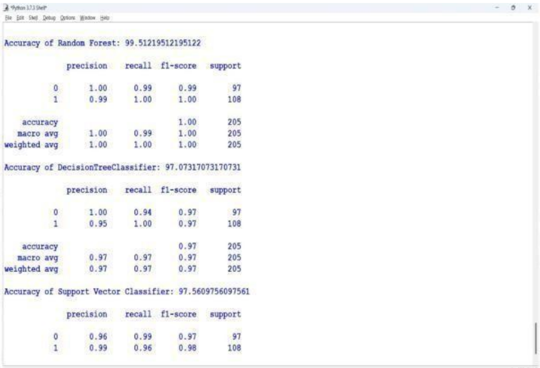


Fig (c): Algorithm Accuracy Report

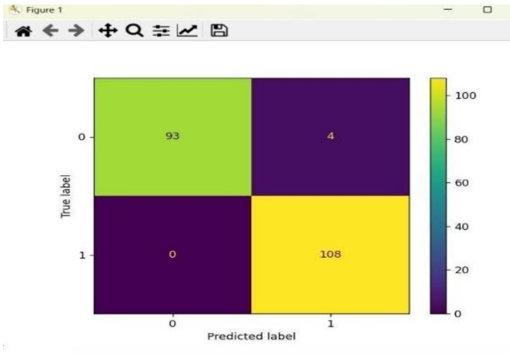


Fig (g) Confix matrix

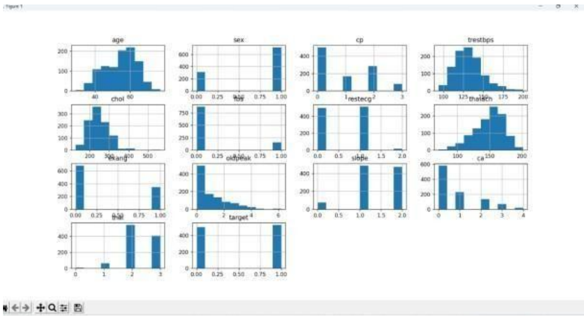


Fig (d): Heart Disease Report

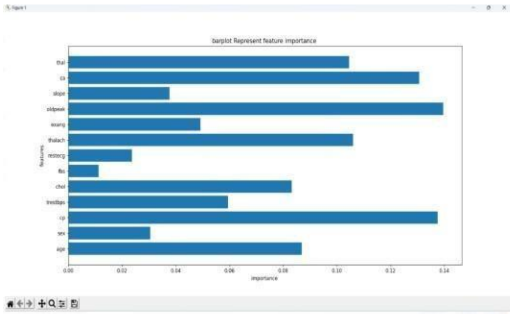


Fig (h) barplot Algorithm Chart

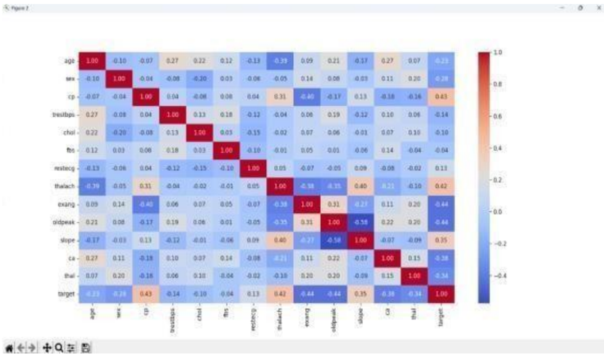


Fig (e): Head Map

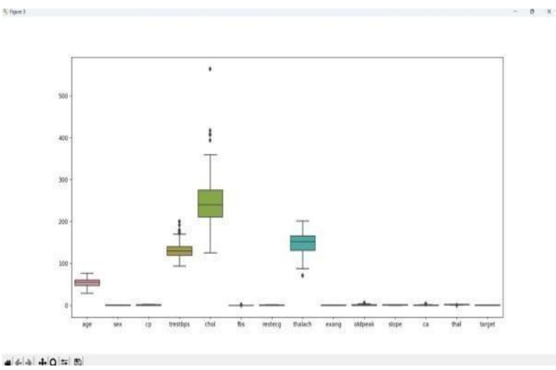


Fig (f) Overall Chart

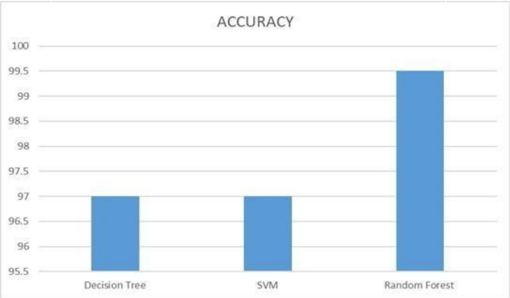
Performance Analysis

The word Heart disease is used for the blood vessel system & heart disease in it. Several factors have been discovered that increase the risk of Heart disease: Family history, Smoking, Unhealthy diet, High blood pressure, High blood cholesterol, Obesity, Physical inactivity, Hypertension, etc., Factors such as these are employed to study Heart disease. In most instances, diagnosis is mostly the patient's present test result & physician's experience. The diagnosis is therefore a compound task that requires a lot of experience & high level of skill.

Table 4.2 Performance Analysis

Model	Accuracy	Score
Decision Tree	97	97.07
SVM	97	97.56
Random Forest	99.51	0.99

Chart 4.2 Performance Analysis



Conclusion

Approximately 18 million individuals 7% of Indians suffer from heart disease. Heart disease is most common in the 65-above age group. This paper primarily discusses three various classes of heart diseases namely coronary artery disease, cardiovascular disease, and cardiomyopathy. The overall goal of this work is to predict solely the existence of heart disease. In this thesis, additional input attributes of portliness and smoking are taken to achieve additional precise results. Heart disease is a frequent cause of death all over the world, and earlier detection results in an improved clinical outcome for patients. In this study, Random Forest (RF) and Support Vector Machine (SVM) have been utilized to forecast heart disease risk based on patients' past medical records, this research highlights the potential for revolutionizing the diagnosis of heart disease while enhancing patient care as well as creating more effective healthcare systems. In the coming years, these predictive models will be integral aspects in driving personalized medicine and population health outcomes by way of regular updates in technology and data.

Acknowledgement

This paper was presented at the Two-Day National Level Seminar on “Realising Sustainable Development Goals in the Indian Context” held at S.E.A. College of Science, Commerce and Arts (Autonomous) during 5th – 6th March 2026.

Reference

Gonsalves, A. H., Thabtah, F., Singh, G., & Mohammad, R. M. A. (2019). Prediction of Coronary Heart Disease using Machine Learning: An Experimental Analysis. ICDLT 2019, Xiamen, China. Association for Computing Machinery.

Limbitote, M., Damkondwar, K., Mahajan, D., & Patil, P. (2020). A Survey on Prediction Techniques of Heart Disease using Machine Learning. International Journal of Engineering Research & Technology (IJERT), 9(6).

Molla, S., Shamrat, F. M. J. M., Rafi, R. I., Umaima, U., Arif, M. A. I., & Hossain, S. (2022). A predictive analysis framework of heart disease using machine learning approaches. Bulletin of Electrical Engineering and Informatics, 11(5), 2705- 2716

Ramya Perumal, Kaladevi AC (2020). Early Prediction of Coronary Heart Disease from Cleveland Dataset using Machine Learning Techniques. International Journal of Advanced Science and Technology, Vol. 29, No. 6, pp. 4225 - 4234.

Saladi Novya Sree, K. Balaji Pranav Reddy, Bangarulakshmi Mahanthi, Dr. S. S. Nandini (2023). An Analysis of Heart Disease Prediction Using Machine Learning. International Journal for Multidisciplinary Research, Volume 5, Issue 3.

Nagaraj M. Lutimath, Neha Sharma, Byregowda B. K (2021). Prediction of Heart Disease using Biomedical Data through Machine Learning Techniques. EAI Endorsed Transactions on Pervasive Health and Technology, Vol. 7, Issue 29. doi: 10.4108/eai.30-8-2021.170881.

Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access, Vol. 7, pp. 81542-81554.

Jian Ping Li, Amin Ul Haq, Salah Ud Din, Jalaluddin Khan, Asif Khan, Abdus Saboor (2020). Heart Disease Identification Method Using Machine Learning Classification in E- Healthcare. IEEE Access, Vol. 8, pp. 107562107574.

V.V. Ramalingam, Ayantan Dandapath, M Karthik Raja (2018). Heart Disease Prediction Using Machine Learning Techniques: A Survey. International Journal of Engineering & Technology, Vol. 7, No. 2.8, pp. 684-687.

Amya Perumal, Kaladevi AC (2020). Early Prediction of Coronary Heart Disease from Cleveland Dataset using Machine Learning Techniques. International Journal of Advanced Science and Technology, Vol. 29, No. 6, pp. 4225-4234.

Saladi Novya Sree, K. Balaji Pranav Reddy, Bangarulakshmi Mahanthi, Dr. S. S. Nandini (2023). An Analysis of Heart Disease Prediction Using Machine Learning. International Journal for Multidisciplinary Research, Volume 5, Issue 3.

Nagaraj M. Lutimath, Neha Sharma, Byregowda B. K (2021).

Prediction of Heart Disease using Biomedical Data through Machine Learning Techniques. EAI Endorsed 10.4108/eai.30-8-2021.170881.

Amya Perumal, Kaladevi AC (2020). Early Prediction of Coronary Heart Disease from Cleveland Dataset using Machine Learning Techniques. International Journal of Advanced Science and Technology, Vol. 29, No. 6, pp. 4225-4234.

Saladi Novya Sree, K. Balaji Pranav Reddy, Bangarulakshmi Mahanthi, Dr. S. S. Nandini (2023). An Analysis of Heart Disease Prediction Using Machine Learning. International Journal for Multidisciplinary Research, Volume 5, Issue 3.

Nagaraj M. Lutimath, Neha Sharma, Byregowda B. K (2021). Prediction of Heart Disease using Biomedical Data through Machine Learning Techniques. EAI Endorsed Transactions on Pervasive Health and Technology, Vol. 7, Issue 29. doi: 10.4108/eai.30-8-2021.170881.

Amya Perumal, Kaladevi AC (2020). Early Prediction of Coronary Heart Disease from Cleveland Dataset using Machine Learning Techniques. International Journal of Advanced Science and Technology, Vol. 29, No. 6, pp. 4225- 4234.

Saladi Novya Sree, K. Balaji Pranav Reddy, Bangarulakshmi Mahanthi, Dr. S. S. Nandini (2023). An Analysis of Heart Disease Prediction Using Machine Learning. International Journal for Multidisciplinary Research, Volume 5, Issue 3.

Nagaraj M. Lutimath, Neha Sharma, Byregowda B. K (2021). Prediction of Heart Disease using Biomedical Data.