CARDIOVASCULAR DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES

J. Nageswara Rao¹, Sahithi Masimukku², Jignas Dagadappula², Sai Chaitanya Devireddy², Kiran Kocherla²
¹Sr. Asst. Professor, Department of Computer Science and Engineering,
²UG Students, Department of Computer Science and Engineering,
¹,²Lakireddy Bali Reddy College of Engineering, AP, India, 521230.

ABSTRACT
The heart is the most important organ of the human body. Heart disease, also known as cardiovascular disease, refers to a variety of diseases that affect the heart. In recent decades, several types of heart disease have often led to death. There are many risk factors for this type of heart disease, so more time is needed to develop a more reliable, effective and stable early diagnosis method. It is important to predict and diagnose heart disease patients. We have also created a diagnostic method using coronary artery disease predictors and cardiac data sets to help us predict diseases based on various information. The final results show that Random Forest produces more reliable results compared to various ML techniques.

Index Terms: Heart Disease, Machine Learning, Random Forest Classifier, SVM, K-NN

I. INTRODUCTION
Coronary heart disease (HD) may be the most confusing and dangerous disease in the world. In this case, the heart is usually unable to pump out the amount of blood needed for other parts of the body to perform its functions normally, leading to heart failure. In the past decade, heart disease or coronary artery disease has become the leading cause of death worldwide. The World Health Organization (WHO) approximately ensures that 17.8 million people die from heart disease each year, and 80% of the main deaths are from coronary heart disease. Heart disease, such as heart failure, heart disease, coronary blood pressure, heartburn, congenital heart disease, heart disease, lung disease, and cardiovascular disease, is a variety of heart diseases that can easily cause. This is an important and accurate diagnosis and treatment of early-onset heart disease. It plays a very important role and helps to take preventive and safety measures.

II. LITERATURE REVIEW
Coronary artery disease afflicts thousands of citizens around the world and remains the main cause of death. Therefore, it is important to detect the disease early to avoid death. It is expected that the disease will be used to develop successful medical diagnostic tools. To minimize the cost of successful screening procedures, medical diagnosis should be professional, safe and computer-assisted. This section provides an in-depth review of related topics, such as machine learning and various machine learning techniques, as well as concise descriptions, preliminary data analysis, various algorithms used to predict diseases.

Heart diseases can be categorized as cardiomyopathy and cardiovascular disorders. Coronary artery disease (CAD) is a major subgroup of cardiovascular diseases and causes severe illness, disability, and even death by reducing the blood and oxygen supply to the heart muscles [1].

The first signs of heart diseases include dyspnea on exertion, palpitation, a sense of numbness or a pain in the center of the chest, and dizzy spells or fainting fits [2]. Given the fatal nature of heart diseases, it is crucial to discover the reasons behind such disorders. Indeed, accurate diagnosis of cardiac abnormalities has been the major goal of many scientific endeavours in the past few decades. A great deal of information is usually collected in the examinations of CAD patients, and the processing of such information can reveal the relationship between the main features of cardiac disorders (e.g. blood pressure, amount of cholesterol, etc.) and the probability of the occurrence of these disorders.
Lee et al. [4] used HRV features for the diagnosis of CAD and applied several algorithms like CPAR, CMAR, SVM, and C 4.5 to them and expressed the performance in terms of TP, FP, Precision, and Recall. The SVM algorithm exhibited the best performance.

Srinivas et al [5] probed into the reasons behind the clogging of each of the major heart vessels, namely the left circumflex artery, left anterior descending coronary artery, and right coronary artery, and achieved an accuracy rate of 84%.

Kavitha et al. [6] utilized the neural network and genetic algorithms for the diagnosis of CAD and succeeded in elevating the accuracy of this method.

Wang et al. [7] employed the Framingham function and features such as sex, age, cholesterol, high-density lipoprotein cholesterol (HDL), blood pressure, diabetes, and smoking for the diagnosis of CAD.

Palaniappan et al. [8] drew upon a limited number of the above-mentioned features in conjunction with three different data mining algorithms, i.e. the decision tree, Naïve Bayes, and neural network, and attained an 85.53% accuracy rate.

Shantakumar et al.[1,2,3][9] used 13 features and neural network algorithm in order to design a system for predicting heart attack.

III. APPROACHES

**Machine Learning:**

Machine learning (ML), which is a part of Artificial Intelligence (AI) to study computer algorithms that develop automatically using information and data. Development machines create a model using sample data, known as "training data". Internal and vascular diagnostic for the patient is a difficult task and requires a variety of details, laboratory tests, and reports to be analyzed. This project does not eliminate the traditional diagnostic method but predicts the risk of heart failure, research has tried to support this procedure using new and modern methods such as Machine learning. Machine learning is used to diagnose diseases in the health care industry. In this study, we used a method using multiple algorithms in the cardiovascular database to determine the correct solution, which could increase the accuracy of predictability for people with heart disease and healthy people.

**Supervised Learning:**

Intargeted supervised learning, the algorithm learns from labeled databases, which provides an answer key algorithm so that it can be used to test its accuracy in training data based on the input-output mapping pair. In this supervised study, the data is divided into two separate sections of the train database and the test database. Whenever we train a model with a database and check the function of the database to determine the accuracy of the model. Supervised learning can be used for all setbacks and divisions.

**Unsupervised Learning:**

In unsupervised learning, the machine is trained using separate or unencrypted data and allows the algorithm to work on that data without any guidance. The task of the machine here is to collect random data according to similarities, patterns, and variations without previous data training. Unlike supervised learning, for unsupervised learning, there will not be any guidance no training will be provided by the machine. Therefore, the machine is restricted from accessing hidden structures to the data without the label itself.

**Reinforcement Learning:**

Reinforcement Learning is one of the three AI models. This model neither utilizes the marked dataset nor the outcomes that are related to information, hence the model gains from the previous experience. In this method, the model upgrades its outcomes dependent on its relationship with the environment and sorts out some ways to examine its issues, and gets the correct result via evaluation and testing different determinations.

**Classification Techniques in Machine Learning:**

Classification is the process of grouping things according to their similar features. This technique predicts the accuracies based on the previous information or the datasets that are given as a training model. There are various classification algorithms for this prediction. In this exploration paper, we utilized the factual algorithms of Machine Learning, for example, Decision Tree, Naïve Bayes, SVM, KNN, Random Forest Classifier, and so on.
By examining and analysis of these various algorithms the coronary illness can be anticipated early and expeditiously.

**Data Source:**
Here for the study, we have taken the UCI repository Machine Learning Cleveland Dataset which consists of 13 different attributes such as Chestpain, Blood pressure, Cholesterol, heart rate, etc [1, 2, 3]. In this paper, we have used 5 different algorithms to predict the disease with the highest possible accuracies.

**Data pre-processing:**
Processing of the dataset is to get accurate predictions from the data set. The dataset is handled with missing data or noisy data if any. The cleaning of the dataset is then done and the dataset is then analyzed with various machine learning algorithms and the accuracies are predicted. Here the below figure shows the process of identifying and handling data.

<table>
<thead>
<tr>
<th>S.No</th>
<th>Attribute</th>
<th>Representative Icon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>Age</td>
</tr>
<tr>
<td>2</td>
<td>Sex</td>
<td>Sex</td>
</tr>
<tr>
<td>3</td>
<td>Fasting blood sugar</td>
<td>Fbs</td>
</tr>
<tr>
<td>4</td>
<td>Chest pain</td>
<td>Cp</td>
</tr>
<tr>
<td>5</td>
<td>ST depression</td>
<td>Oldpeak</td>
</tr>
<tr>
<td>6</td>
<td>Rest blood pressure</td>
<td>Trestbps</td>
</tr>
<tr>
<td>7</td>
<td>Maximum Heart rate</td>
<td>Thalch</td>
</tr>
<tr>
<td>8</td>
<td>Serum cholesterol</td>
<td>Chol</td>
</tr>
<tr>
<td>9</td>
<td>Rest electrocardiograph</td>
<td>Restecg</td>
</tr>
<tr>
<td>10</td>
<td>Exercise-induced angina</td>
<td>Exang</td>
</tr>
<tr>
<td>11</td>
<td>Thalassemia</td>
<td>Thal</td>
</tr>
<tr>
<td>12</td>
<td>Slope</td>
<td>Slope</td>
</tr>
<tr>
<td>13</td>
<td>No. of vessels</td>
<td>Ca</td>
</tr>
</tbody>
</table>

Table1 Attributes of the heart data set
Various Algorithms

**K-Nearest Neighbor:**

The K-Nearest Neighbor is perhaps the most effortless approach to adapt precisely dependent on a coordinated learning measure. The K-NN calculation takes the similarity between new data and accessible datasets and places the new data point in the class that intently resembles the available classifications. The K-NN algorithm stores all accessible data and isolates the new data point as per the similitude. This implies that where new data arises it very well may be handily classified into a valuable application utilizing the KNN. KNN algorithm can be used for regression and partitioning but predominantly for partition problems.

**Naïve Bayes:**

Naïve Bayes algorithm is another supervised learning that comes under a machine learning algorithm that is based on the Bayes theorem which helps to solve classification problems. It is widely used in text classification which incorporates high-quality training databases, algorithms that help to build faster learning machine models that can make faster predictions. It is a potential arrangement, which implies it predicts the premise of the likelihood of something. Some mainstream instances of the Naïve Bayes Algorithm are spam sifting, Sentimental investigation, and article characterization.

\[
P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)}
\]

Where,

P(A|B) is the posterior probability, P(B|A) is the likelihood probability, P(A) is the prior probability, P(B) is the marginal probability.

Naive Bayes is a straightforward, easy to execute, also proficient characterization algorithm that handles non-direct, convoluted information. Notwithstanding, there exists a deficiency of precision as it depends on presumption and class contingent independence. It obtained an accuracy of 85.25% is achieved using the Cleveland dataset.

**Decision Tree:**
The Decision tree is a very sturdy and popular algorithm for division and prediction. The decision tree is a tree-like structure, where each internal node defines a test in a qualification, each branch addresses a test outcome, and every terminal node holds a category label that works with category details and numbers. The decision tree is used to build structures such as trees. The decision tree is straightforward and broadly utilized for managing clinical databases. It is simple to use and to interpret the data in a tree-shaped structure. The algorithm partitions the data into two or more identical sets according to the most important indicators. For each characteristic, the entropy is determined and the details are categorized, with indicators for higher or minimum entropy information:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

$$Gain(S, A) = H(S) - \sum_{S \in S} \frac{|S|}{|S|} H(S)$$

The outcomes acquired are simpler to peruse and analyze. The Decision tree algorithm has higher exactness in contrast with different algorithms as it dissects the dataset in the tree-shaped diagram. Notwithstanding, the data might be excessively grouped and just every trait is tested in turn for decision making. It has obtained an accuracy of 81.97%.

Random Forest:
A random forest algorithm is a supervised machine learning algorithm. In this algorithm, several trees form a forest. Every tree in this forest meets the expectations of the category and the category with the most votes becomes the model prediction. In random forest planning, the count of trees provides high accuracy. It is utilized for partitioning and retrieval function, but can do well with partition function, and can overcome lost values. Besides, the delay in getting predictions as it needs huge datasets and many trees. It obtained 90.16% exactness using the Cleveland dataset.

Support Vector Machine:
The Support Vector Machine is another basic calculation that each machine learning expert ought to have in their stockpile. In this case, the support vector machine is liked by numerous individuals as it creates high precision with low estimation power. However, it is broadly used for division purposes. The purpose of a vector support machine algorithm is to identify a hyperplane in the N-dimension space (N - number of elements) that separates the data points.

Logistic Regression:
Logistic regression is another famous Machine Learning algorithm. It is utilized for depicting the categorical dependent variable by using a set of independent variables. It gives probabilistic outcomes which lie somewhere in the range of 0 and 1. Logistic Regression is similar to Linear Regression yet Logistic regression is utilized for solving the problems of Classification. In Logistic regression, we fit an “S” molded strategic capacity, rather than fitting a line of regression, which anticipates the two most extreme values 0 or 1.

IV. RESULTS
The Output of this proposed framework will give an outcome if the individual has a coronary illness, as far as 1 or 0. Various machine learning algorithms are used for this study which gives the accuracy of the persons having a heart disease.

The system gives an idea about the heart status which can lead to Coronary Artery Disease (CAD). If a person is having heart disease the results obtained will be 1 and otherwise, the result will be 0. Based on our analysis the Random forest Classifier has the highest accuracy of all as in the table below. If there is a positive output, then the patient should approach a cardiologist for a diagnosis. The results of the accuracy of the algorithms obtained during the data test are shown in the following table.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-Nearest Neighbor (KNN)</td>
<td>67.21%</td>
</tr>
</tbody>
</table>
In this paper, we proposed prediction of heart disease system using various ML techniques. The supervised learning algorithms are applied to the same data set in which random forest has obtained the highest accuracy for the model for which the disease is predicted with 0 or 1. In conclusion, as pointed out, we believe only the limited success achieved in creating a hypothetical model for the patients who are having heart disease and so integrated and complex models must predict the early onset of heart disease with increasing accuracy. For future scope using various electronic devices build with android in which patients can easily check their status about their disease. In future we enhanced advanced novel techniques for heart disease prediction.

REFERENCES:

7. Jaymin Patel, Prof.TejallUpadhyay, Dr. Samir Patel,"Heart Disease Prediction Using Machine learning and Data Mining Technique".
9. V.V. Ramalingam, Ayantandapath, M Kartik Raja," Heart disease prediction using machine learning techniques: a survey".
10. Fahd Saleh Alotaibi," Implementation of Machine Learning Model to Predict Heart Failure Disease".