Predicting Cardiovascular Disease with Machine Learning

Mr.C.Rajeshkumar¹, Dr.K.Ruba Soundar², Dr.M.Vargheese³, Dr.G.Nallasivan⁴, A.Selvakumar⁵

¹Assistant Professor/CSE,PSN College of Engineering and Technology,Tirunelveli,India,  
²Associate Professor/CSE,Mepco Schlenk Engineering College,Sivakasi,India,  
³Professor/CSE, PSN College of Engineering and Technology,Tirunelveli,India,email:  
⁴Professor/CSE, PSN College of Engineering and Technology,Tirunelveli,India,email:  
⁵CSE Student, PSN College of Engineering and Technology, Tirunelveli,India,

Abstract: Machine learning has many applications in today's modern environment. In the medical field, there are no exceptions. Predicting the presence of problems in locomotion, cardiac function, and other areas may be greatly aided by machine learning. Such information, if predicted in advance, may provide vital intuitions to physicians, allowing them to tailor their diagnosis and treatment plan to each individual patient. Here, patient records are read from comma-separated values (CSV) file. After obtaining the data, the procedure is executed, and a valuable heart attack level is produced. We're training machine learning algorithms to identify those at risk for getting heart disease. Classifiers such as decision trees, logistic regression, support vector machines, and random forests are compared and contrasted in this research.

Keywords: Machine learning, Heart Disease Prediction, Decision tree algorithm, Random Forest, Support Vector Machine, Logistic Regression.

INTRODUCTION

According to the World Health Organization, around 12 million people worldwide lose their lives each year due to cardiovascular disease. Heart disease is a prominent cause of death and disability across the world. Predicted cardiovascular disease is one of the most important issues in the field of data analysis. Recent years have seen a dramatic increase in the global incidence of cardiovascular disease. Since death from heart disease often occurs suddenly and without warning, it has earned the nickname "silent killer." An early diagnosis of heart disease is critical for helping high-risk patients make decisions about whether or not to make lifestyle changes that reduce the severity of the condition.

In this case, the use of machine learning techniques may be quite useful. Despite the fact that heart disease may appear in a variety of ways, there is a common set of fundamental risk factors that determine whether or not someone will eventually be at risk for heart disease. We may say that this method is well
suited to accomplishing heart disease prediction by collecting data from various sources, classifying them under suitable categories, and then analyzing to acquire the required data.

**Motivation**

The major purpose of this research was to suggest a model for predicting the onset of cardiovascular disease. In addition, this study's objective is to zero in on the best approach to heart disease categorization. Supporting this work is a comparative study and evaluation of four classification algorithms: SVM, Decision Tree, Logistic Regression, and Random Forest. Despite the widespread use of machine learning techniques, accurately forecasting heart illness remains a mission critical challenge. As a result, the three algorithms are evaluated using a wider range of criteria and testing procedures. Because of this, researchers and doctors will be able to develop a more effective.

**LITERATURE SURVEY**

Recent years have seen a flurry of noteworthy articles detailing the results of trials and research into the intersection of medical science and machine learning.

Using hill climbing and decision tree algorithms, Purushottam et al. suggested a "Efficient Heart Disease Prediction System" in their study [1]. They took the Cleveland dataset, and before using classification methods, they preprocessed the data. To complete the Knowledge Extraction, an open-source data mining method called Evolutionary Learning (KEEL) is used to infer the missing variables. A decision tree is built from the top down. One node is chosen at each stage of the hill climbing algorithm based on the result of a test. The used parameters and values may be trusted. It has a confidence interval of at least 0.25. The method has an approximate 86.7% rate of accuracy.

Prediction of cardiac disease using machine learning algorithms was suggested by Santhana Krishnan et al. in [2], namely the use of a decision tree and the Naive Bayes method. In a decision tree algorithm, branches are constructed according to criteria that determine whether an node is True or False. Algorithms like support vector machine and k-nearest neighbor used depend on factor to determine whether to divide the data vertically or horizontally. In contrast, a decision tree is a tree-like structure in which each node, or branch, is based on a previous choice. Using a decision tree may also provide light on which data points are most crucial. Cleveland has also been utilized in their analysis. Some approaches divide the dataset such that 70% is used for training and 30% is used for testing. The accuracy of this method is 91%. Naive Bayes, another classification method, comes in at number two. Because it is capable of dealing with complex, nonlinear, dependant data, it is deemed appropriate for the heart disease dataset. The algorithm achieves an accuracy of 87%.

A neural network approach called multilayer perceptron is used for both training and evaluating the dataset in the study "Prediction of Heart Disease Using Machine Learning" presented by Aditi Gavhane et al. There will be one input layer and one output layer in this algorithm, with one or more hidden levels in between. Each input node is linked to the output layer through hidden layers. Some arbitrary weights are put on this link. The connection between the nodes may be feedforward or feedback, and the other input is termed bias, which is assigned with weight depending on necessity.

Avinash Golande et al. published "Heart Disease Prediction Using Effective Machine Learning Techniques" in [4], which used a limited set of data mining methods to aid physicians in making distinctions between
various forms of heart disease. K-nearest neighbor, decision tree, and Naive Bayes are the most often used methods. Packing calculation, Part thickness, consecutivenegligible streamlining and neural systems, straight Kernel self-arranging guidance, and SVM(Boltzmann Machine) are some more examples of characterization-based techniques that are used.

More risk factors for cardiovascular disease are taken into account in the "Machine Learning Techniques for Heart Disease Prediction" suggested by Lakshmana Rao et al. in [5]. Thus, distinguishing cardiac illness is challenging. Various neural networks and data mining methods are utilized to assess the severity of heart disease in individuals.

In [6], Abhay Kishore et al. introduced a heart attack prediction system that employs Deep learning methods and makes use of a Recurrent Neural System in order to forecast the likely elements of the patient's heart-related illnesses. In order to provide the most accurate model with the fewest possible errors, this model employs deep learning and data mining. This works serves as a gold standard against which other heart attack prediction models may be judged.

The primary goal of the "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" proposal by Senthil Kumar Mohan et al. in [7] is to increase precision in predicting cardiovascular diseases. KNN, LR, SVM, and NN are the algorithms used in the heart disease prediction model with hybrid random forest with linear model (HRFLM), which results in an enhanced exhibition level with a precision level of 88.7 percent.

Performance of prediction for two categorization models is studied and compared to prior work as proposed by Anjan N. Repaka et al. in [8]. The experimental findings demonstrate that our suggested strategy outperforms competing models in terms of accurately predicting the proportion at risk.

4. METHODOLOGY
To get a feel for how far ML has come in its application to heart disease prediction, we combed through the existing literature. Finding research gaps and providing direction for further studies in a topic are two key benefits of doing a comprehensive literature review. All relevant papers are retrieved from online resources, synthesized, and presented utilizing a method in SLR to answer the research challenges described in the study. Newcomers to the field may get insight into the state of the art thanks to the new perspectives and knowledge provided by an SLR study.

Dataset Collection:
We begin by amassing a dataset to serve as the backbone of our heart disease prediction algorithm. After collecting data, we split it into a training and a testing set. The training dataset is used to teach the prediction model, while the testing dataset is used to assess its performance. In this project, we only put 70% of the data through its training paces and 30% through its testing paces. The project's dataset came from the UCI Heart Disease Database.

Selection of attributes:
Attribute or feature selection encompasses the process of deciding which characteristics will be used in the prediction system. The purpose of this is to improve the efficiency of the system. Multiple patient factors, such as gender, the kind of chest pain, fasting blood pressure, serum cholesterol, etc., are selected for the
prediction. In this approach, a correlation matrix is used to decide which attributes to include.

**Pre-processing of Data:**

One of the most important steps in creating a machine learning model is the pre-processing of data. Inaccurate findings might be the consequence of poorly cleaned or formatted data being fed into the model. In order to get the information we need, we must first "pre-process" it. It is used to deal with the ambiguity, redundancy, and missing values included in the dataset. Data pre-processing includes tasks like importing datasets, splitting datasets, scaling attributes, etc. Improving the model's precision requires pre-processing the data.

**Balancing of Data:**

It is possible to balance an unbalanced dataset in two ways. Both problems—under and oversampling—are present.

**UnderSampling**

Dataset balance in UnderSampling is achieved by reducing the size of the large class. When there is enough data, this procedure is taken into account.

**OverSampling**

When oversampling, the tiny samples are enlarged to restore data balance. This method is considered when there is insufficient data. When we are itching to dive headfirst into studying a massive dataset and building an ML model. We keep getting a "out of memory" issue whenever we try to load the dataset into our computer. When dealing with massive datasets, this happens all the time. When it comes to data science, one of the biggest obstacles is handling "big datasets" on computationally limited devices (a problem that may be solved, of course, with more resources). The term "sampling" refers to a statistical technique used to manage a bigger dataset.

**Prediction of Disease**

For classification, a variety of machine learning algorithms are employed, including SVM, Decision Tree, Random Tree, Logistic Regression, Ada-boost, and XG-boost. Algorithms are compared, and the one that predicts heart disease with the best degree of accuracy is chosen.

<table>
<thead>
<tr>
<th>diabetes</th>
<th>totChol</th>
<th>sysBP</th>
<th>diaBP</th>
<th>BMI</th>
<th>heartRate</th>
<th>glucose</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>195</td>
<td>106</td>
<td>70</td>
<td>26.97</td>
<td>80</td>
<td>77</td>
</tr>
<tr>
<td>0</td>
<td>250</td>
<td>121</td>
<td>81</td>
<td>28.73</td>
<td>95</td>
<td>76</td>
</tr>
<tr>
<td>0</td>
<td>245</td>
<td>127.5</td>
<td>80</td>
<td>25.34</td>
<td>75</td>
<td>70</td>
</tr>
<tr>
<td>0</td>
<td>225</td>
<td>150</td>
<td>95</td>
<td>28.58</td>
<td>65</td>
<td>103</td>
</tr>
</tbody>
</table>

Table: 1 Heart disease dataset

**Step 1:** Loading the csv Dataset

**Step 2:** Plotting Histogram and creating a heatmap

**Step 3:** Performing Smoting processes to classify the Positive and Negative Results.

**Significance:** The smoting results are concerned with the four algorithms.
EXPERIMENT AND ANALYSIS

Accuracy calculations and project execution for Heart Disease are complete. Four algorithms are used to specify the user’s heart rate, blood pressure, glucose level, smoking status, and body mass index, with the best result expected to be Highest Accuracy. The algorithms carried out in the following stages:

The results are accurate, as you can see here. Data from the Heart Disease Prediction Model are read into the input dataset.
The above figures show the Area Under ROC Curve for the Logistic Regression, K-Nearest Neighbour, Decision Trees, Support Vector Machine.

The above figure illustrates the Accuracy of Dataset.
The above figure illustrates the F1 score mean of the dataset.

CONCLUSION

As one of the main causes of mortality in India and throughout the globe, heart disorders would have a profound societal effect if promising technologies like machine learning were used to the early prediction of heart issues. Significant medical progress may be made with the early diagnosis of heart illness, which may aid high-risk individuals in deciding on preventative lifestyle changes. Heart disease is becoming more prevalent among the general population. Because of this, prompt identification and treatment are essential. Using the right kind of technological aid in this area might be very beneficial for both doctors and patients. SVM, Decision Tree, Random Forest, and Logistic Regression are just some of the seven machine learning methods put to the test. The dataset was subjected to both Adaptive Boosting and Extreme Gradient Boosting.

The 76 feature dataset represents the expected factors that lead to heart disease in people, and 14 relevant factors are selected from it for evaluation purposes. Taking into consideration all of the system's characteristics results in less efficiency for the author. The goal of attribute selection is to boost productivity. Here, n features are selected for use in evaluating the correctness of different models. Because of their high associations with other variables, some properties of the dataset are discounted. If every attribute in the dataset is used, performance suffers dramatically.

REFERENCE


