# HCI RECOMMENDATIONS FOR TREATMENT DECISIONS

**Vikrant Chaudhary**
Amity Institute of Information Technology, Amity University, Noida-125, UP, India
vikrantc8218@gmail.com

**Dr Deepa Gupta**
Amity Institute of Information Technology, Amity University Noida-125, UP, India
dgupta@gmail.com

*Abstract—*
*A stimulating field of research that has attracted a lot of attention recently is heart disease prediction using ML technology. This study's objective is to generate and explore and test machine learning (ML) techniques that can reliably and precisely forecast a person's risk of covering heart disease based on a number of clinical and non-clinical factors. Large amounts of medical data, including patient demographics, medical histories, lifestyle choices, and genetic data, are analyzed to create these models. Disease prediction using ML has a number of advantages, such as earlier diagnosis, more individualized care, and better patient outcomes. However, there are a number of difficulties with this research as well, considering the need for top-quality data, privacy issues, and ethical issues with the use of personal data. The proposed system evaluates the user-provided symptoms as input and outputs the likelihood of the condition. It implements different machine learning models to evaluate their performance on predicting diseases. In addition, comparison between all the algorithms is also done to determine the best ML algorithm for the prediction of diseases.*

*Keywords – Disease prediction, Heart disease, KNN, SVM*

## I. INTRODUCTION

Myocardial Infraction, also known as heart attack was once considered to be an issue for the elderly. For someone to experience a heart attack under the age of 40 was extremely uncommon back in the days but now it can be seen in

Heart disease prediction using machine learning (ML) is an significant research area in medical science. Heart disease is a dominant source of mortality in the world, and pre stage diagnosis and accurate forecasting of heart disease can significantly improve outcomes of the patient.

Machine learning (ML) has shown significant potential in disease prediction by analysing generous volumes of medical data and learning the patterns and relationships that can help predict the onset or progression of various diseases. Here are some of the ways in which ML can be used for disease prediction:

1. Early detection: ML algorithms can analyze patient data and identify early warning signs of diseases such as cancer, heart disease, and diabetes. This can enable doctors to intervene early and provide effective treatment before the disease progresses.
2. Risk assessment: ML models can analyze patient data and identify risk factors that increase the likelihood of developing specific diseases. This can assist medical professionals in creating individualised disease prevention plans.
3. Treatment optimization: ML can analyze patient data and identify which treatments are most effective for different diseases, helping doctors optimize treatment plans for each patient.

The initial step in utilizing machine learning to forecast cardiac disease is to gather and preprocess the pertinent data, such as demographics, medical history, symptoms, and test outcome. Selection of feature and feature engineering approach can be applied to narrow down the dataset's dimensions and determine its most crucial features.

The following step is to divide the data into training and testing datasets after it has been pre-processed. The dataset for testing is used to assess the performance of the model after the ML model has been trained using the training dataset.

Common ML algorithms used for heart disease prediction include decision trees, random forests, neural networks, support vector machines, and logistic regression. Various metrics for evaluation can be used to gauge how well something ML model, including accuracy, precision, recall, and F1 score. The ML model can be fine-tuned by adjusting the hyper parameters of the algorithm to optimize the model's performance.

In summary, heart disease prediction using ML is a promising approach to enhance the effectiveness and precision of cardiac disease detection and prognosis. By leveraging large volumes data of the patient & applying ML algorithms, healthcare professionals can develop accurate predictive models for heart disease, which can help in early detection, timely intervention, and improved patient outcomes.

## II. REVIEW OF LITERATURE

Disease prediction is an essential area of research in healthcare, and machine learning (ML) has shown great potential in this field. This literature review summarizes recent studies that use ML to predict various diseases.

- ML algorithms were employed in a study by Islam et al. (2020) to forecast Parkinson's disease based on voice recordings. They used (SVM) to gauge the risk of Parkinson's illness and (CNN) to excerpt features from the speech recordings. Their algorithm demonstrated the potential of ML for early identification of with a 95.2% accuracy rate of Parkinson's disease.

- Another research by Yang et al. (2021) used ML algorithms to forecast prostate cancer revolve on magnetic resonance imaging (MRI) scans. They developed a deep learning algorithm that could forecast the existence of prostate cancer with great accuracy. They also used explainability techniques to pinpoint the geographic areas of MRI scan that were most informative for the forecast.

A significant global public health concern is heart disease, particularly early identification of risk factors can significantly reduce mortality rates. The use of Machine learning (ML) algorithms has been prevalent in medical research, including heart disease prediction. This literature review summarizes recent studies that utilize ML to predict heart disease.

- One study by Attia et al. (2019) used ML algorithms to estimate the risk of acquiring heart disease using electrocardiogram (ECG) data. They developed an algorithm that outperformed traditional risk factors such as age, sex, and smoking history. Another research by Khened et al. (2020) used ML algorithms to forecast heart disease revolving on clinical and demographic data. They compared several ML models and found that a decision tree algorithm had the relevantly high accuracy.

- Another research by Nguyen et al. (2020) used a deep learning algorithm to predict heart disease using ECG signals. They used a long short-term memory (LSTM) network to forecast the outcomes and a convolutional neural network (CNN) to extract characteristics from the ECG signals likelihood of heart disease. Their algorithm demonstrated the promise of deep learning for heart disease prediction, achieving an accuracy of 94.5%.

- A research by Bhandary et al. (2020) used ML algorithms to estimate heart disease risk revolving on patient data from electronic health records. They used a logistic regression algorithm and a random forest algorithm to foretell the risk of heart disease, achieving an accuracy of 84.6% and 85.9%, respectively. They also identified factors of risk such as age, diabetes, and high blood pressure that were strongly associated with heart disease.

## III. FRAMEWORK

In this study, a web application is developed that can predict heart disease. High level algorithms are first tested to predict the disease. The algorithms with the best accuracy and efficiency is then selected and integrated into the website.

The web portal takes in multiple input fields as symptoms from the end users who wish to check if have the disease or not. The system then will take the symptoms and use it in the machine learning model.

*A. Machine learning Model*

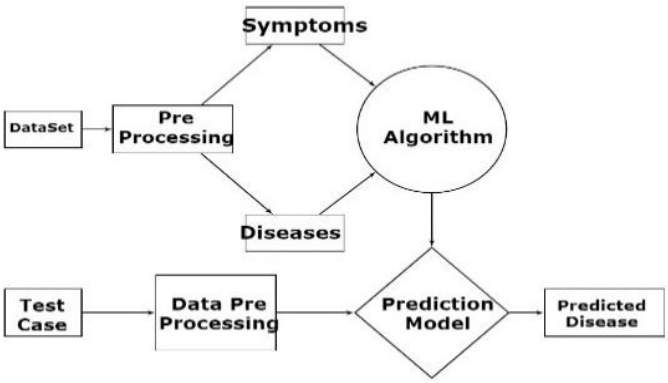


**Fig. 1**

*B. User Interface Model*

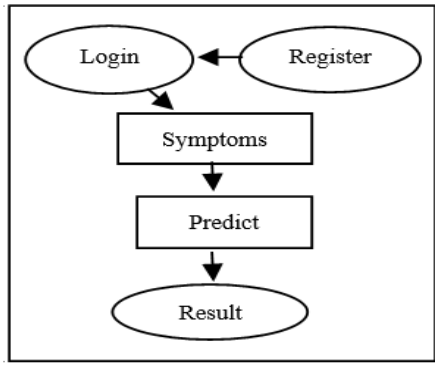


**Fig. 2**

## IV. MACHINE LEARNING

Before starting the analysis, let's look into the process flow of machine learning.
Examples of typical workflow processes include data collection, preprocessing, creating datasets, model training and improvement, evaluation, and deployment to production.

*A. Data Gathering (Dataset)*

The University of CMLR's Cleveland data was the source of initial information.
https://archive.ics.uci.edu/ml/datasets/heart+Disease

It's also accessible in a modified form on Kaggle.
https://www.kaggle.com/ronitf/heart-disease-uci

>Df.head()

| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 63 | 1 | 3 | 145 | 233 | 1 | 0 | 150 | 0 | 2.3 | 0 | 0 | 1 |
| 1 | 37 | 1 | 2 | 130 | 250 | 0 | 1 | 187 | 0 | 3.5 | 0 | 0 | 2 |
| 2 | 41 | 0 | 1 | 130 | 204 | 0 | 0 | 172 | 0 | 1.4 | 2 | 0 | 2 |
| 3 | 56 | 1 | 1 | 120 | 236 | 0 | 1 | 178 | 0 | 0.8 | 2 | 0 | 2 |
| 4 | 57 | 0 | 0 | 120 | 354 | 0 | 1 | 163 | 1 | 0.6 | 2 | 0 | 2 |

**Fig. 1 First 5 values of Dataset**

### *B. Attributes*

The table consists of 1025 rows × 14 columns. All the attributes are listed in table 1.

Table 1 Attributes

| |
|---|
| Age |
| gender |
| CP |
| Trestbps |
| Cholestrol |
| Fbs |
| Restcg |
| Exang |
| Old peak |
| slop |
| CA |
| Thall |

### *C. Classification*

A fundamental task in machine learning (ML) is classification, which entails categorizing or labelling every observation based on its attributes. It is widely utilized in many different applications, including sentiment analysis, fraud detection, image and speech recognition, and many others. In this article, we will dive into the world of classification algorithms with Python and explore some of the most popular and powerful models.

- Naïve Bayes
- Logistic Regression(LR)
- Random Forest Classifier
- Decision Trees Classifier
- K-nearest Neighbors
- SVM

### *D. Preprocessing*

Preprocessing data helps to clean and prepare the data for subsequent analysis, which is why it is a significant stage in data analysis and machine learning. The following are some typical data pre-processing steps:

- Cleaning: In this process, incorrect or missing data is removed or imputted.
- Transformation: In this stage, the data's format is changed to one that is more suitable for analysis.
- Feature Selection: In this step, the analysis's most significant features or variables are chosen.

1. Univariate Analysis: analysing each variable separately. Plotting graphs and figures to understand the behaviour of each attribute
2. Bivariate Analysis: Studying the relation among independent and its dependent numerical features.

- Discretization: In this step, continuously collected data are converted into discrete values.
- Normalization: In this step, the data are normalized to eliminate any potential bias.

### *E. Algorithms*

The following algorithms are used in this study to determine the most accurate ML model for predicting Heart Diseases. Before analyzing and comparing the models, studying them is necessary.

### *A.* Logistic Regression

In binary classification issues where the output variable (dependent variable) can only take one of two values. LR is a linear method in which a logistic function is used to model the connection between the input variables (independent variables) and the output variable. Any real-valued input is translated into a value between 0 and 1 using the logistic function, also referred to as the sigmoid function.

Algorithm:
Using the maximum likelihood estimation (MLE) algorithm, the logistic regression model is constructed by estimating the coefficients of the linear equation that best fits the data. The equation is of the form:

$$\log(\text{odds}) = Y0 + Y1x1 + Y2x2 + ... + Ynxn$$

Model evaluation
By measuring metrics such as accuracy, precision or area under the receiver operating characteristic (ROC) curve using the testing set, the logistic regression model's performance is assessed.

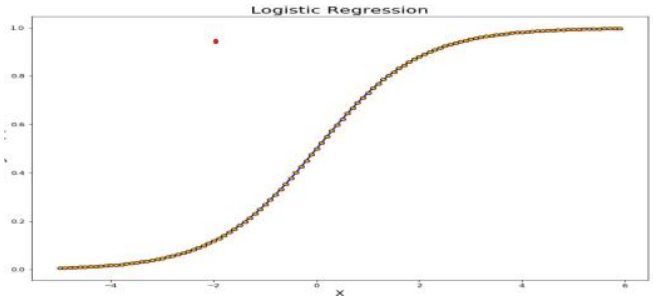


**Fig 3. Logistic Regression**

*B.* K- Nearest Neighbors

A straightforward but efficient machine learning method called K-Nearest Neighbor (k-NN) is utilized for two classification and regression problems. It is an instance-based learning technique in which the algorithm memorizes the training dataset rather than building a broad internal model.

A new observation's class or value can be predicted using k-NN by using the "k" nearby training observations as a starting point. The user-defined option "k" indicates how many nearest neighbors should be taken into account. Although there are different distance metrics that can be utilized, the Euclidean distance metric is commonly used to calculate the distance between two observations.

Algorithm:

- Load the data set, which includes the predictor and response variables.
- Choose the number of neighbours you want: Pick the Kth neighbour (K) whose information will be used for classification.
- Pick a distance measurement: To determine how close the data points are to one another, choose a distance metre.
- Determine the separation between each observation's KN neighbours.
- Using a voting mechanism to predict the response variable, using the variable that received the most votes in the testing set.
- Calculating the accuracy, recall, and F1 score on the testing set will allow you to assess the model's performance.
- Repeat steps 6 through 8 for different K values, and then choose the K value that provides the best performance on Use the trained KNN model to predict outcomes for fresh data points.

*C.* Decision Tree

A well-liked and commonly used machine learning approach for both classification and regression tasks is the decision tree. It is a kind of supervised learning technique that creates a model in the shape of a tree structure to make judgements or predictions in response to a set of input features.

The tree structure is made up of nodes, which stand in for features or attributes, branches, which represent the feature's potential values, and leaves, which stand in for the result or choice. The algorithm maximises information gain or minimises data impurity by iteratively dividing the data into subsets based on the most crucial attribute.

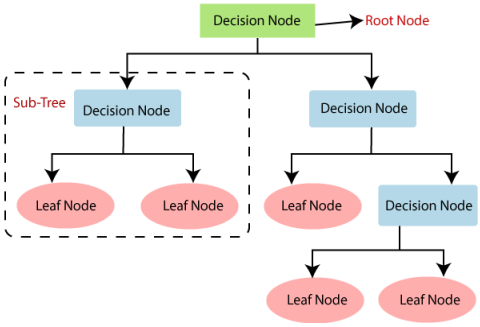


**Fig.4 Decision Tree**

*D.* Random Forest

An ensemble learning technique called random forest is used for both classification and regression applications. It consists of a number of decision trees, everyone of which was built using a random portion of the data training and a random subset of the input features.

Building various decision trees on different subsets of the data training and infusing their forecasting to generate a last prediction is how the method operates. In classification tasks, by considering the majority vote of all the trees' projections, the final prediction is determined. The final prediction in regression tasks is calculated by averaging the forecasting of all the trees.

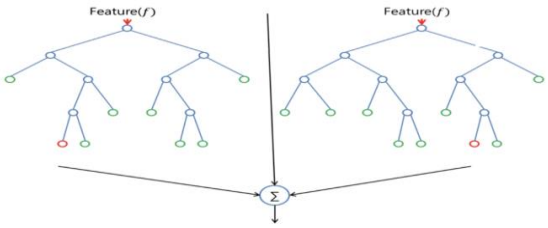


**Fig.5 Random Forest**

Working algorithm:

- Random sampling: At each iteration of the algorithm, a random sample of the training data is selected for building a decision tree.
- Feature selection: Only a random subset of the features is taken into account for splitting the data at each decision tree split.
- Tree building: Recursively dividing the data into the desired feature sets creates the decision tree.

- Ensemble learning: Following the construction of each decision tree, the predictions made by each are pooled to create the final result.
- Performance evaluation: Metrics including accuracy, precision, recall, F1 score, and mean squared error (MSE) are calculated using the testing set to assess the random forest model's performance.

### *E.* Support Vector Machine

A support vector machine (SVM) is an approach to supervised learning algorithm used in regression and classification. SVMs are particularly useful in cases where the data has multiple dimensions, making analyzing and visualizing is challenging.

Finding the hyperplane that best convert the data into different classes is the significant tenet of SVMs. The margin, also known as the distance between the nearest points of each class, is maximised by selecting the hyperplane in this manner. This ensures that the classifier has the best possible generalization performance on unseen data.

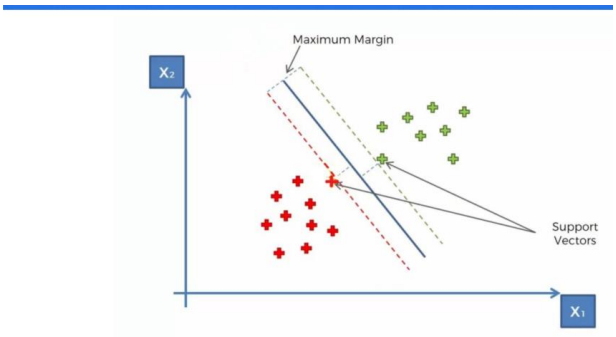


**Fig.6 SVM**

Model evaluation: By determining metrics like accuracy, precision, recall, F1 score, or area under the ROC curve and applying them to the testing set, it is possible to assess how well the SVM model performs.

## V. ANALYSIS AND DISCUSSION

After determining and contrasting the precision of each machine learning model, performance is optimal for the decision tree algorithm. It has the highest accuracy of 92%.

Table 2 Accuracy Table

| ALGORITHM | ACCURACY |
|---|---|
| **Random Forest** | **86.440678** |
| **Support Vector Machine** | **83.050847** |
| **Naive Bayes** | **81.355932** |
| **Logistic Regression** | **79.661017** |
| **Decision Tree** | **76.271186** |

Since Random forest algorithm has the best performance in forecasting heart diseases, this study will integrate Random forest algorithm into the web development platform.
Now let's check whether the person have chances of heart disease or not using some random values. Fig.5 shows the results of the model

```
y_predict = dt_model.predict([[1,148,72,35,
79.799,33.6,0.627,50]])
print(y_predict)
              if y_predict==1:
print("Heart Patient")
else:
print("Not a Heart Patient")
```

Result:
>Heart Patient

## VI. SNAPSHOTS

A) Full page screenshot of the website

A single page, fully responsive website is shown in fig.7 which is designed using MERN Stack development.
MERN stack involves MongoDB, express, reactJs, NodeJs.


**Fig.7 Full page Screenshot**

B) About Us page

Introduction of the developer as well as the motive behind this application is stated in the about us page which can be seen in Fig 7.

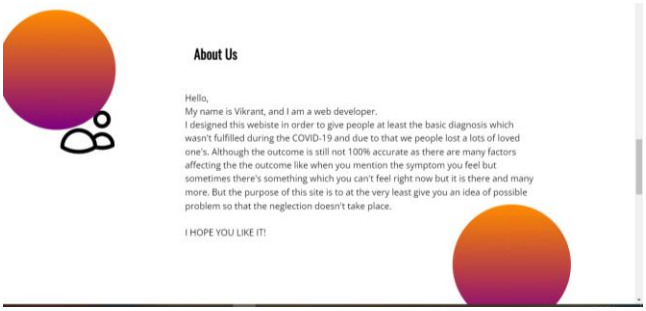


**Fig.8 about us**

C) Prediction page

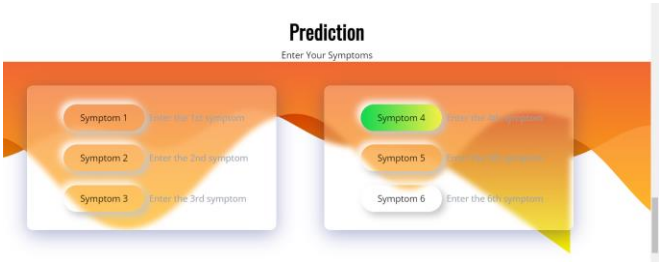


**Fig.9 Prediction page**

D) Results page

The figure below (fig 10) is a separate page displaying the results calculated by the algorithm. It tell whether a person suffers from Heart Disease or not


**Fig.10 Result Page**

## VII. CONCLUSIONS

In this research paper, we first looked at the concepts of Machine learning and its applications in the healthcare industry. We also researched the various machine learning methods employed in the medical field. In the realm of disease prediction, machine learning's capacity to sift through enormous volumes of data and find patterns that may be challenging for

humans to comprehend has shown great promise. Although ML has the potential to completely transform disease prevention and detection, it must be utilized ethically and with an awareness of its ethical implications. This paper concludes that among different high level algorithms, decision tree algorithm has proved to be the most accurate of all.

## REFERENCES

[1] Garg, S., & Sharma, P. (2019). Predicting heart disease using machine learning algorithms. International Journal of Engineering and Advanced Technology, 8(5), 546-551.

[2] Mirza, M. W., & Alahmadi, T. (2021). Heart disease prediction using machine learning techniques: a systematic review. International Journal of Medical Informatics, 146, 104355.

[3] Rehman, S., & Raza, M. Q. (2020). A comparative analysis of machine learning techniques for heart disease prediction. Journal of Medical Systems, 44(4), 82.

[4] Hameed, W., Zafar, F., & Abbas, H. (2021). Disease of heart prediction using machine learning models: a systematic review. Computer Methods and Programs in Biomedicine, 204, 106062.

[5] Kumar, A., & Maitra, M. (2021). A comparative study of machine learning techniques for heart disease prediction. Journal of Ambient Intelligence and Humanized Computing, 12(11), 12979-12995.