Vol 7, Issue 1, 2024 **Impact Factor: 5. 355** DOI: https://doi.org/10.31426/ijamsr.2024.7.1.7012

IJAMSR

**International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281**

# A Novel Analysis of Heart Disease Prediction Through Machine Learning Algorithms

**Deepak Upadhyay**

**Ankit Garg**

*deepakupadhayay7779@gmail.com*
*snankitgarg@gmail.com*

**ABSTRACT**

Heart disease remains the primary cause of execution globally, driving the adoption of advanced medical technologies for treatment. However, a significant challenge in medical centers is the variability in expertise among healthcare professionals, leading to suboptimal decisions and adverse outcomes. Machine learning algorithms and data mining techniques offer a solution by enabling predictive diagnosis and facilitating automated assessment in hospitals. By analyzing various health parameters, including age, sex, cerebral palsy (CP), blood pressure (BP), and fasting blood sugar (FBS) levels, heart disease prediction becomes feasible. This study explores the efficacy of different algorithms, including Statistical Regression, Naive Bayes Analysis, KNN, Decision Tree, SVM, XGBoost, Neural Network, and Random Forest, utilizing a built-in dataset. Our research compares the accuracy of these techniques, revealing Random Forest to achieve the highest accuracy of 90.16%.

***Key Words: Naive Bayes, k Nearest Neighbour (KNN), Decision-tree model, XGBoost, Support Vector Machine (SVM), Artificial neural network (ANN), Random Forest, Logistic Regression, Heart Disease.***

## 1. INTRODUCTION

Heart disease is among the most prevalent causes of death worldwide, surpassing all other causes with an estimated annual toll of 12 million deaths. In the United States alone, heart disease claims a life every 34 seconds, often resulting from blocked blood flow to the heart or brain, leading to heart attacks. People at risk of heart disease may exhibit elevated blood pressure, glucose, and cholesterol levels, along with stress, all detectable through basic health assessments. Cardiovascular disease (CVD) leads to significant illness, disability, and mortality. The diagnosis of heart disease poses a complex and crucial challenge in medicine. While medical diagnosis is essential, it can be daunting, especially considering the scarcity of specialist physicians and resources in certain areas. Data mining is an effective method for revealing hidden patterns and insights that improve decision-

IJAMSR

**International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281**

making processes. This approach proves pivotal with healthcare providers in delivering reliable diagnoses and superiority care. Providing support for healthcare professionals lacking specialized knowledge and skills is imperative. A notable limitation of existing methods lies in their capacity to consume precise findings when required. Our goal method predicts heart disease through methods for data mining and machine learning algorithms such as Logistic Regression, Naïve Bayes, k Nearest Neighbour (KNN), a decision tree, the Support Vector Machine (SVM), XGBoost, a neural network, and Random Forest (RF), based on specific health parameters.

## 2. RELATED WORK

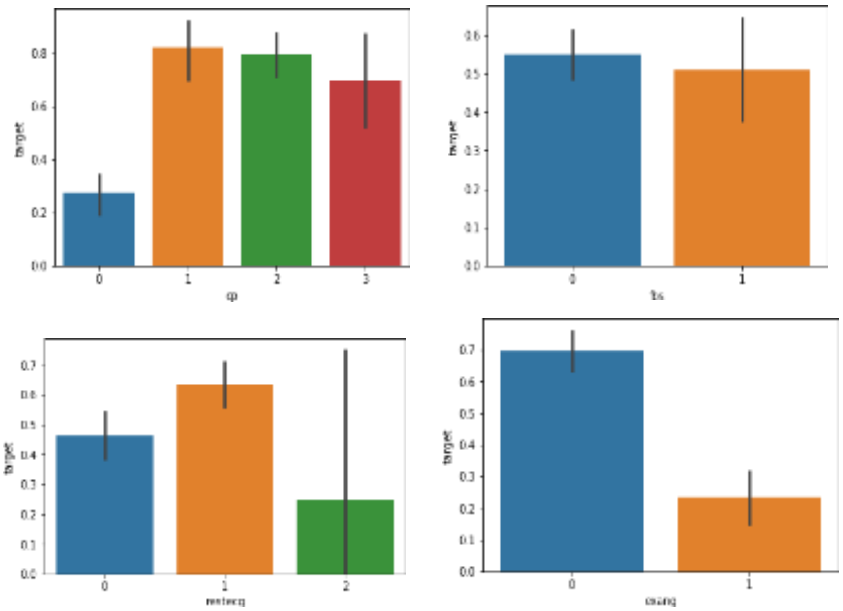
[1] Utilizing data from the UCI repository, heart disease prediction is proposed using KStar, J48, SMO, and Bayes Net, alongside Multilayer Perceptron with WEKA software. Despite achieving satisfactory accuracies, SMO (89%) and Bayes Net (87%) outperform other techniques, yet overall performance remains unsatisfactory.[2]In another study, Kaggle data is employed to predict stroke patients using Neural Network and Support Vector Machine (SVM), achieving accuracies of 81.97% and 81.97% for XGBoost and Decision tree respectively in the training dataset, and 83.61% and 81.97% in the test dataset.[3]Assesses various machine learning algorithms using UCI repository data, with Random Forest received the maximum precision of 90.16%. followed by Naïve Bayes and KNN at approximately 85.15% and 67.21%, and Decision Tree at 81.97%. [4] The WEKA tool is used to evaluate machine learning algorithms, where ANN with PCA initially achieved 94.5% accuracy, rising to 97.7% post-PCA application, indicating a significant improvement.[5] Furthermore, in cardiovascular disease prediction, Random Forest achieved the highest accuracy of 90.16%. Although Artificial Neural Network (ANN) demonstrated a slightly lower accuracy of 81.97% compared to other models, it was chosen as the final model to ensure a balance between precision and transparency in predicting heart disease. [6] Despite higher accuracy rates in some models, the artificial neural network (ANN) with a lower accuracy of 84.25% was chosen to maintain a balance between precision and transparency in predicting heart disease. [7] Additionally, the Hidden Naïve Bayes algorithm achieved 100% accuracy, surpassing traditional Naïve Bayes in predicting heart disease.

## 3. METHODOLOGY

The fundamental objective of our suggested approach is to allow for early identification of cardiac disease by accurately forecasting its onset. Our solution uses multiple data gathering approaches and artificial intelligence algorithms, including the following: Naive Bayes, KNN, Decision Tree, and ANN and Random Forest. These algorithms use health data to forecast the possibility of heart disease. For data analysis, we use Navigator's notebook system Jupyter is a free to use platform. that allows you to create numerous machine-learning methods using library imports. Additionally, Anaconda Navigator allows us to download necessary libraries via Anaconda Prompt.

IJAMSR

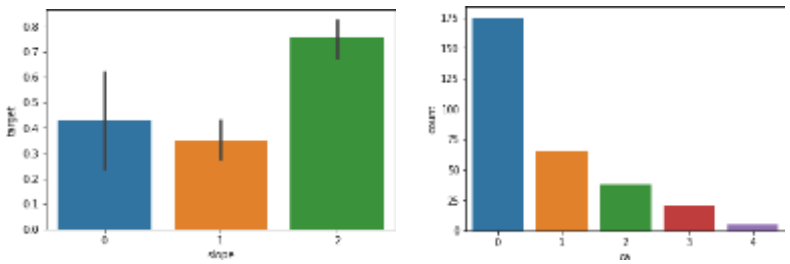
**International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281**



Its interactive environment enables live code execution, visualization creation, data processing, and graph plotting, enhancing the efficiency of our analysis and modeling process.

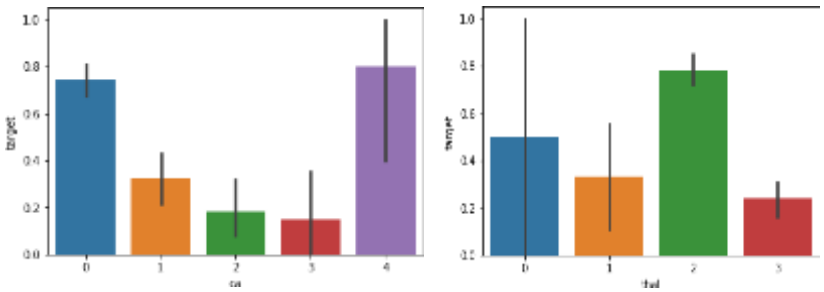
## 4. DATA SELECTION

We used an included database of the repository for machine learning at UCI to predict cardiac disease. This dataset has fourteen attributes:



Age Sex

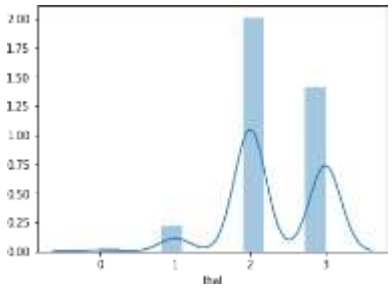
Cerebral palsy or chest pain (CP) Blood Pressure (BP) in millimeters Hg Cholesterol (mg)



Fasting Blood Sugar Test (FBS) Resting electrocardiographic findings Thalach (maximum heart rate reached). Exang (Exercise-induced angina) Old peak (ST depression caused by exercise compared to rest)

IJAMSR

**International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281**



Slope (Slope of the peak exercise ST section) CA (Number of major vessels).

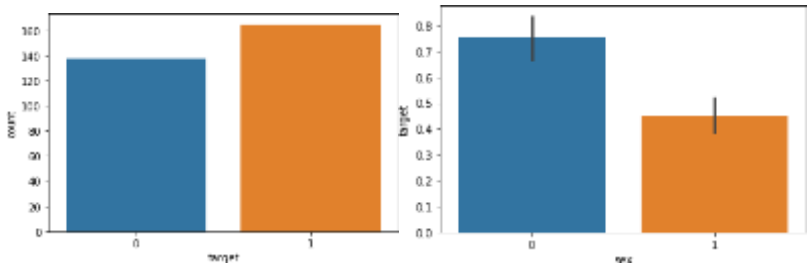
Thal (Reversible Defect) Target: (0 or 1)

These attributes collectively inform predictive modeling for heart disease.

## 5. DATA SPLITTING

The set of data is divided into testing and training sets, with 25% of the data allocated for testing and 75% for training. Prior to model training, data normalization is conducted to handle NaN (missing) values.

## 6. VISUALIZING TRAINING AND TESTING DATA SPLIT AND DATA NORMALIZATION



The accuracy and efficacy of each study are evaluated using Common measures include the True Positive (TP) rate, True Negative (TN) rate, recall, accuracy, and the F-measure.

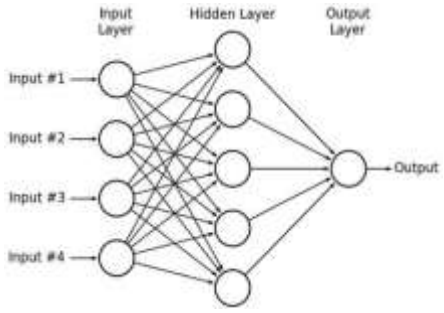
## 7. ALGORITHMS USED FOR EXPERIMENTS

### Artificial Neural Network

Artificial neural networks replicate the structure of human neurons, featuring interconnected nodes linked by directional connections. Each node serves as a processing unit, with the connections representing causal relationships. In clinical decision-making, Artificial Neural Networks aid doctors in efficient and accurate analysis and decision-making processes. Typically, an artificial neural network initiates with an input layer, when every device connects to nodes in hidden layers, which may then connect to produce multiple layers. This categorization technique is gaining recognized as a strong tool for data mining, serving various purposes in descriptive and predictive data analysis. Refer to the figure below for a sample representation of an artificial neural network.

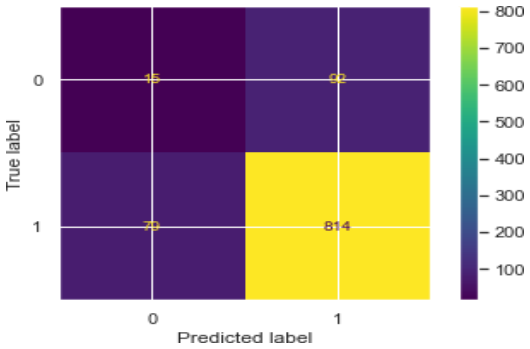
IJAMSR

**International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281**




AI for Neural Networks represent essential parts of computer networks meant to emulate, analyze, and Utilize data akin to the human brain. Possessing self-learning capabilities, they continually improve results with increased data availability.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.750000 | 0.6 | 0.666667 | 5.0 |
| 1 | 0.666667 | 0.8 | 0.727273 | 5.0 |
| accuracy | 0.700000 | 0.7 | 0.700000 | 0.7 |
| macro avg | 0.708333 | 0.7 | 0.696970 | 10.0 |
| weighted avg | 0.708333 | 0.7 | 0.696970 | 10.0 |




**k Nearest Neighbor (KNN)**

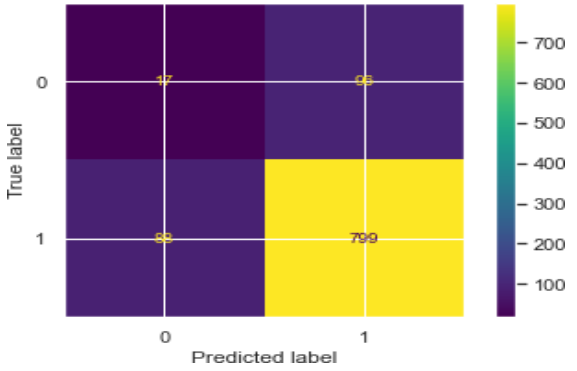
KNN, a commonly used machine learning algorithm, is favored for scenarios with continuous parameters. It operates by predicting the nearest neighbor for classification, known for its simplicity and speed compared to other algorithms. KNN is versatile and able to manage both regression and classification employment. Through leveraging the heart disease dataset, KNN determines whether an individual has heart disease by calculating distances between data points on a graph. Our implementation of KNN involves classifying individuals based on factors like age and sex. Unlike some models, KNN does not require training data for model development because it uses trained data during the testing process. It saves all cases and categorizes the latest information based on its proximity to the nearest neighbors. KNN has two critical stages: identifying the k quantity of representations in the data collection and using these instances to find the nearest neighbor. The correct score obtained with KNN is 67.21%.

IJAMSR

**International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281**

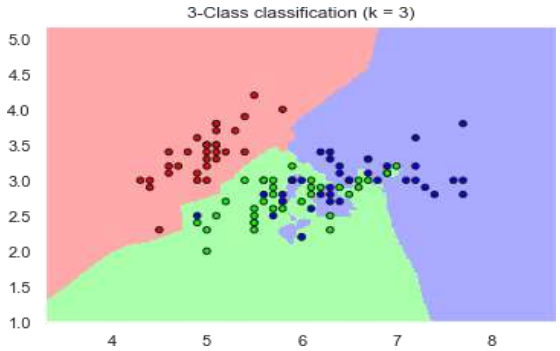
### Confusion Metrics

Confusion metrics are used to assess the accuracy of a KNN is dependent by the length of the measure and the value of K.




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.666667 | 1.000000 | 0.8 | 4.0 |
| 1 | 1.000000 | 0.666667 | 0.8 | 6.0 |
| accuracy | 0.800000 | 0.800000 | 0.8 | 0.8 |
| macro avg | 0.833333 | 0.833333 | 0.8 | 10.0 |
| weighted avg | 0.866667 | 0.800000 | 0.8 | 10.0 |

### Visualization




The graph illustrates patients with heart disease depicted by red dots, while those without heart disease are represented by blue dots.

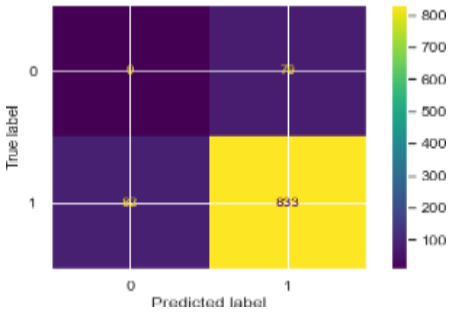
### Decision Tree

A decision tree is a supervised training conduct which classification that is well-known for its simplicity and interpretability. It is capable of processing both numerical and category data successfully. A decision tree is structured like a tree, with foundation, segment, and leaf nodes in between. All branches reflect the values of a given characteristic in the dataset, whereas internal nodes test attributes and leaf nodes represent the projected class or outcome. The classification process in a decision tree begins at the root node and advances to the leaf nodes using predicted features and specified rules. CARTthe ID3, C4.5 as a J48, which is and the CHAID algorithm are among the most used decision tree algorithms for sickness prediction.

IJAMSR

**International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281**

### Confusion Metrics




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.500 | 0.500 | 0.500 | 6.0 |
| 1 | 0.250 | 0.250 | 0.250 | 4.0 |
| accuracy | 0.400 | 0.400 | 0.400 | 0.4 |
| macro avg | 0.375 | 0.375 | 0.375 | 10.0 |
| weighted avg | 0.400 | 0.400 | 0.400 | 10.0 |

The accuracy score achieved using Decision Tree is: 90.16 %

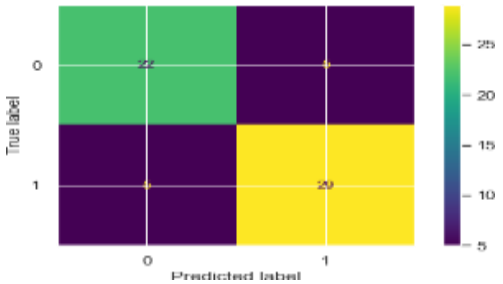
### XGBoost

XGBoost, or Extreme Gradient Boosting is a robust and accurate data mining algorithm that excels in both classification and regression problems. It is part of the ensemble learning method, specifically the boosting family, which combines numerous weak learners to produce a strong learner. laxity with various regularization strategies while handling missing values and delivering feature priority scores to improve interpretability. XGBoost's optimized data structures and support for distributed computing result in great training speed and memory efficiency, making it appropriate for large-scale datasets across multiple platforms and programming languages.

### Confusion Matrix

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.5 | 0.6 | 0.545455 | 5.0 |
| 1 | 0.5 | 0.4 | 0.444444 | 5.0 |
| accuracy | 0.5 | 0.5 | 0.500000 | 0.5 |
| macro avg | 0.5 | 0.5 | 0.494949 | 10.0 |
| weighted avg | 0.5 | 0.5 | 0.494949 | 10.0 |

The accuracy score achieved using XGBoost is: 83.61 %




## 8. DISCUSSION

Our study focuses on the use of information mining techniques for medical purposes, specifically regarding cardiac analysis. Trials on a dataset about heart illness, using five different data mining algorithms. Our goal was to find the best effective algorithm for predicting heart disease and to see which algorithm produced the highest accuracy. We carried out five trials with the same goal, comparing the performance of KNN, decision trees, neural networks, decision trees, naive bayes, and random forests.

## 9. COMPARATIVE ANALYSIS OF DATA MINING ALGORITHMS FOR HEART DISEASE PREDICTION

The study we conducted sought to employ artificially intelligent algorithms to diagnose heart issues in healthcare. To accomplish this, we tested multiple algorithms on patients with heart disease. Through implementing these algorithms, we hoped to find the best classification algorithm for predicting heart disease.

IJAMSR

**International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281**

| Algorithms | Accuracy | TN | FP | FN | TP |
|---|---|---|---|---|---|
| ANN | 0.81 | 42 | 11 | 5 | 30 |
| KNN | 0.67 | 35 | 08 | 3 | 26 |
| Decision Tree | 0.81 | 42 | 11 | 5 | 30 |
| XGBoost | 0.83 | 44 | 12 | 5 | 32 |
| Random Forest | 0.90 | 21 | 6 | 5 | 29 |

Following the execution of various algorithms, the next stage was to compare their performance. We aimed to select the algorithm with the maximum accuracy. To help this comparison, we used a variety of performance indicators, including the accuracy statistic, True Positive, False Positive, False Negative, True Negative, and the ROC Curve. Below is a summary table showcasing the performance of the algorithms used in our experiments. Upon examining additional performance metrics, including True Positive (TP) rates, it is immediately apparent that KNN has the highest TP rate (40), whereas Decision Tree has the lowest (29). Conversely, when considering False Positive (FP) rates, Decision Tree records the highest at 15, while Artificial Neural Network (ANN) displays the lowest at 2. Comparatively, Naïve Bayes, KNN, and ANN demonstrate similar accuracies, with all exhibiting favorable TP rates. Notably, KNN dominates with an FP rate of 7, ahead of naive Bayes prediction at 4 and ANN with 2. Given the sensitivity and gravity of heart disease, which accounts for millions of deaths annually, it is imperative to prioritize high TP rates and low FP rates. Accurate and timely disease diagnosis significantly impacts patient outcomes, underscoring the importance of algorithmic performance.

## 10. CONCLUSION

Our research focuses on the implementation of data mining technologies in medical care, specifically the identification of heart illness, which is potentially lethal. We evaluated the performance of various algorithms, such as KNN as the model deep neural networks, a decision tree, Naive Bayesian, and Random Forest, with defines such as Accuracy, True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP). We reported the findings of five tests on the same dataset in tabular format for better understanding and comparison. Our data show that Naive Bayes has the highest accuracy (88%), followed by ANN and KNN (87%). This highlights the advantages of data analysis for medical purposes for early disease detection and diagnosis. In terms of future study, we recommend conducting additional research to improve classification accuracy using sophisticated techniques such as Bagging and Support Vector Machine. Furthermore, it is critical to evaluate the Evaluate the efficiency of all algorithms and employ the solution suggested to the suitable healthcare sector. To improve accuracy, additional features can be incorporated into the algorithms, and stakeholders should leverage this tool to make informed decisions. While our implementation did not involve parameter tuning, future iterations could benefit from adjustments to optimize performance. Furthermore, future research should look into the utilization of larger datasets connected to heart attack and various findings decline. approaches to improve estimation outcomes. Ultimately, utilizing high-quality datasets free from inconsistencies could lead to more accurate and reliable predictions of heart disease.

IJAMSR

**International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281**

## REFERENCES

1. C. Beyene, P. Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", 118(8):165-173 · January 2018.
2. Muhammad usama riaz, shahid mehmood awan, abdul ghaffar khan, "prediction of heart disease using Artificial neural network", October 2018
3. Ujma Ansari, Jyoti Soni, Dipesh Sharma, Sunita Soni. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", March 2011 Data Mining in Healthcare for Heart Diseases.
4. Komal Kumar Napa, G. Sarika Sindhu, D.Krishna Prashanthi, A.Shaeen Sulthana, "Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers", April 2020.
5. Jabbar Akhil, Shirina Samreen, "heart disease prediction system based on hidden naïve Bayes classifier", October 2016.
6. Umair Shafique, Irfan Ul Mustafa, Haseeb Qaiser, Fiaz Majeed, "Data Mining in Healthcare for Heart Diseases", March 2015.
7. Hossam Meshref, "Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach", January 2019.