# Abstract

Heart attack is one of the most pressing problems of the health care industry. In general, the patient's reports must be scrutinized by doctors to make a diagnosis of a heart failure. This research study is an attempt to reduce the efforts and time put in by the doctor by automating the risk prediction with the help of a binary classifier. A prototype implementation of such a system with an easy-to-use user interface is presented in this paper. This study analyzes the Behavioral Risk Factor Surveillance System, survey to test whether self-reported cardiovascular disease rates are higher in Singaram coal mining regions in Andhra Pradesh state, India, compared to other regions after control for other risks. An automated system for medical diagnosis would enhance medical care and reduce costs. In this paper popular Machine Learning techniques namely, Decision Trees, Naïve Bayes and Neural Network are used for prediction of heart disease.

# CHAPTER 1: INTRODUCTION

Acute myocardial infarction, commonly known as a heart attack, remains one of the deadliest cardiovascular events, claiming millions of lives globally each year. This critical condition often occurs when blood flow to the heart is blocked, usually by a clot, leading to severe tissue damage. Early detection and prevention are essential to improving survival rates and reducing the burden on healthcare systems.

In recent years, the advent of machine learning (ML) has revolutionized the field of medical diagnostics. These data-driven techniques excel at analyzing vast amounts of complex data, identifying patterns, and predicting outcomes with remarkable accuracy. Applying machine learning to heart attack prediction allows healthcare providers to identify high-risk patients early, enabling timely intervention and personalized treatment.

This report explores how ML tools like Random Forests, Decision Trees, Neural Networks, and Support Vector Machines (SVMs) are employed to analyze clinical and lifestyle data, enhancing the prediction and prevention of heart attacks. Furthermore, the project integrates various ML techniques with data visualization tools and big data platforms, such as Hadoop, to handle large and complex datasets effectively.

---

## 1.2 Background and Related Work

### 1.2.1 Understanding Heart Diseases: Overview and Diagnosis

Heart disease encompasses a wide array of medical conditions affecting the heart’s ability to function properly. These include coronary artery disease, arrhythmias, and congenital heart defects. The heart, a muscular organ that pumps blood throughout the body, relies on its intricate coronary artery network for oxygen supply. Any disruption in this system can have catastrophic consequences, impacting not just the heart but the entire body.

The traditional diagnostic process for heart-related issues involves clinical evaluations, medical imaging, and pathological testing. While effective, these methods are resource-intensive and can delay early detection. Machine learning offers a promising alternative by automating the analysis of medical data and providing actionable insights.

### 1.2.2 Machine Learning in the Medical Domain

Machine learning is revolutionizing the healthcare industry, becoming an essential tool for hospitals and clinics. With the increasing availability of patient data—ranging from medical histories to diagnostic images and real-time monitoring results—machine learning offers the ability to analyze and interpret this information effectively.

By identifying patterns that may not be immediately obvious, these algorithms are helping healthcare providers make earlier diagnoses, assess patient risks, and tailor treatments to individual needs.

Some practical examples of machine learning in action include:

- Improving the accuracy of diagnoses using image recognition techniques, such as interpreting CT scans or ECG readings.
- Assisting in creating personalized treatment plans, ensuring better outcomes for patients.
- Predicting the likelihood of diseases by analyzing both genetic and environmental factors.

## 1.3 Objectives of the Project

The Heart Attack Prediction Analysis project is designed with the following key objectives:

**Risk Prediction:**
Develop accurate machine learning models to predict the likelihood of heart attacks, enabling early detection and intervention.

**Factor Identification:**
Analyze clinical data and lifestyle patterns to identify the most significant contributors to heart disease.

**Automation in Healthcare:**
Create an automated prediction system to support healthcare providers in making informed decisions for early diagnosis and risk management.

**Addressing Data Challenges:**
Ensure the system maintains data privacy, reduces bias, and delivers fair predictions across diverse patient groups.

**Future Integration:**
Explore innovative applications of advanced technologies, like deep learning and real-time patient monitoring, to improve prediction accuracy and healthcare delivery.

## 1.4 Purpose and Significance

Diagnosing heart attacks is challenging due to the complex interplay of clinical and pathological factors. Timely and accurate predictions can save lives, making heart attack prediction a critical area in healthcare innovation.

The use of ML tools provides unparalleled support in analyzing large datasets and identifying high-risk individuals. By integrating these tools into the diagnostic process, healthcare providers can:

Improve accuracy in identifying at-risk patients.

Reduce the time and cost of diagnosis.

Implement preventative measures based on data-driven insights.

This study seeks to establish machine learning as a reliable and scalable solution to the growing burden of cardiovascular diseases.

---

### 1.5 Technological Landscape

The technological advancements in heart attack prediction span both traditional machine learning models and cutting-edge deep learning frameworks. Below is an overview of commonly used techniques:

Traditional Machine Learning Models:

Logistic Regression: Estimates the probability of heart disease based on predictors like cholesterol and blood pressure.

Decision Trees: Models non-linear relationships and provides interpretability but may overfit.

Random Forests: Combines multiple decision trees for improved accuracy and robustness.

Support Vector Machines (SVMs): Effective for high-dimensional data and binary classification tasks.

k-Nearest Neighbors (k-NN): Predicts outcomes based on the most similar data points.

Advanced Machine Learning and Deep Learning Models:

Gradient Boosting Machines (e.g., XGBoost): Build sequential trees to optimize predictions.

Neural Networks: Deep learning models analyze ECG signals and time-series data.

Convolutional Neural Networks (CNNs): Identify patterns in medical imaging.

Recurrent Neural Networks (RNNs): Capture dependencies in sequential patient records (e.g., time-series ECG data).

Autoencoders: Identify anomalies in heart disease data through unsupervised learning.

---

### 1.6 Scope

The scope of this project extends to developing a robust, scalable, and clinically relevant system for heart attack prediction.

Key Deliverables Include:

Data Preprocessing: Cleaning and preparing clinical datasets.

Feature Selection: Identifying the most relevant predictors of heart disease.

Model Development: Training and evaluating machine learning models.

System Deployment: Creating a prototype system accessible to healthcare professionals.

By building a comprehensive prediction model, this project aims to reduce diagnostic delays and improve patient outcomes. Future extensions could incorporate wearable devices for real-time monitoring and personalized health interventions.

# CHAPTER 2: LITERATURE REVIEW

## 2.1 Literature Review

Heart Disease (HD) HD is defined a range of conditions that affect your heart. It is describing any disorder of the heart. The umbrella of HD consists of different type of HD such as blood vessel diseases (coronary artery disease, and arrhythmias) and heart defects when you are born with congenital heart defects, among others. The term “heart disease” is always used interchangeably with the term “cardiovascular disease (CVD).” CVD generally refers to conditions that involve blocked or narrowed blood vessels that can lead to a heart attack, stroke, or chest pain

Heart Disease Prediction (Chala Beyene, 2018) Proposed a methodology to foretell the occurrence of HD to overcome the problem of diagnosis of HD. It improved the existence methodology by choosing Naïve Bayes, J48, and SVM for predicting the occurrence of HD for early automatic diagnosis in short time to support the qualities of services and reduce costs to save the life of individuals. This methodology uses various attributes of HD to identify whether a patent has HD or not. The comparison of analysis in the dataset is used WEKA software. (P. Sai Chandrasekhar Reddy, 2017) Recommended ANN algorithms for HD prediction system in DM. The main aim of this predicting system is to reduce cost of a diagnosis like different type of test was done to decide for diagnosis of HD.

So, they have proposed a new system to prophesy the condition of the patient based on their parameters such as age, blood pressure, heartbeat rate, cholesterol, etc. and evaluate if a patient has HD or not. The proposed system is provided its accuracy in java. (Dwivedi, 2016) Focused to evaluate the performance of different ML algorithms for HD prediction. The comparison between different algorithms such as Naïve Bayes, KNN, Logistic Regression and Classification tree to identify the high performance for predicting the HD

# CHAPTER 3: PROBLEM STATEMENT

## 3.1 Problem Statement

Heart attack defines a condition that affects a heart. Heart attack contains differences diseases such as coronary artery disease (CAD), Congenital Heart attack, Mitral Value Prolapse, Arrhythmia, Pulmonary Stenosis, Dilated Cardiomyopathy, Heart Failure, Hypertrophic Cardiomyopathy, and Myocardial Infarction. One of them, cardio vascular disease (CVD) is one of the main diseases of the heart that refers to the condition of obstructed blood vessels that can be happened a stroke and heart attack. Another form of HD can be rhythm, heart's muscle, etc. (Mayo Clinic, 2019) CVDs are one of the major cause of people death globally. Many people have died from CVDs compare to other cause. In 2016, due to CVDs, an estimated 17.9 million human died. It is illustrating 31% of human deaths all over the world. Stroke and heart attack have occupied 85% of these deaths. (World Health Organization,

In 2017, the latest fact data of Word Health Organization (WHO) published that Nepal has reached 18.72% or 30,559 deaths from Coronary HD. The rate of age fixed death is 158.35 out of 100,000 population and world rank is #41. (World Life Expectancy, 2019) According to The Heart Foundation; 13% of men and 10% of women are died due to HD in Australia. In 2017, Whilst HD had 18,590 deaths. So that HD was a one four death of cause factor in 2017. (The Heart Foundation., 2019) So, Nepal government also needs to use this system to aware the patient before being critical situation. This system provides accurate result that help to less worry about the doctor’s negligence.

With the consideration of WHO statistical facts, the most powerful causes of death globally are a HD. It seemed to the negligence of patients as well as doctors to increase a HD patient. Some of the difficulties to execute the doctor’s decision and lack of application to clearly diagnosis of HD become the cause of human death. Regarding the above issues, we are proposing a web-based HDPS that is one of the best solutions to efficiently and accurately predict the HD patients. The proposed system eliminates the various testing of HD and supports the decision making of doctors. This system can accept a singleton query and display the clear output of the presence of HD level. This system is useful for any hospital and clinic to evaluate the patient getting a HD. It is reduced the number of tests and provide an efficient output of patient HD.

# CHAPTER 4: IMPLEMENTATION

## 4.1 Introduction and Purpose

Heart attack is one of the leading causes of death worldwide. Timely diagnosis and prediction of heart attack risks can significantly reduce mortality rates by enabling early intervention and preventive care. Traditionally, patient reports must be carefully analyzed by medical professionals to diagnose heart-related conditions, which is both time-intensive and prone to human error.

This study aims to automate the risk prediction process using machine learning techniques to assist doctors in making quicker and more accurate diagnoses. Various algorithms, such as Decision Trees, Naïve Bayes, and Neural Networks, are implemented to build a binary classifier that predicts the likelihood of a heart attack.

Machine learning analytics has become a powerful tool in the healthcare industry for analyzing large datasets to extract valuable insights. By using these technologies, the goal is to build a system capable of mining patient data to predict, manage, and potentially prevent heart attacks. This project demonstrates how machine learning can support medical professionals in delivering efficient and effective patient care while reducing costs.

## 4.2 Data Collection and Preparation

The dataset used for this analysis includes several clinical and lifestyle features relevant to heart attack prediction. Key attributes are:

Age: The age of the patient.

Gender: Male or Female.

Cholesterol Levels: Measured in mg/dL.

Blood Pressure: Resting systolic blood pressure in mm Hg.

Heart Rate: Maximum heart rate achieved.

Chest Pain Type (CP): Categorized as typical angina, atypical angina, non-anginal pain, and asymptomatic.

Fasting Blood Sugar: Whether fasting blood sugar exceeds 120 mg/dL.

Steps in Data Preparation:

Handling Missing Values: Missing values were imputed using mean or median values based on the distribution of each attribute.

Normalization and Scaling: Continuous features like age, cholesterol, and blood pressure were normalized to ensure uniform data distribution.

Encoding Categorical Variables: Variables such as gender and chest pain type were converted into numerical formats using one-hot encoding.

### 4.3 Feature Selection

Effective feature selection is critical to building a robust predictive model. Techniques like Recursive Feature Elimination (RFE) and correlation analysis were used to identify the most significant predictors of heart attacks.

Key Features Identified:

Age

Maximum Heart Rate

Cholesterol Levels

Chest Pain Type

Resting Blood Pressure

These features were selected based on their correlation with the target variable (presence or absence of heart disease).

### 4.4 Machine Learning Models

Three machine learning algorithms were implemented and compared:

1. Decision Trees:

A tree-based model that splits the data into branches based on decision rules. It provides an interpretable model but is prone to overfitting if not pruned correctly.

2. Naïve Bayes:

A probabilistic classifier based on Bayes' theorem, assuming independence among predictors. It is computationally efficient and works well for small datasets.

3. Neural Networks:

A deep learning model that mimics the functioning of the human brain, using multiple layers to capture complex patterns in the data. Neural networks were trained with one hidden layer and ReLU activation for this project.

### 4.5 Model Training and Evaluation

The dataset was split into training (80%) and testing (20%) subsets. Each model was trained on the training data and evaluated on the testing data.

Evaluation Metrics:

Accuracy: Measures the percentage of correctly classified instances.

Precision: The proportion of true positives among all predicted positives.

Recall (Sensitivity): The proportion of true positives identified by the model.

F1-Score: A weighted average of precision and recall.

**Results Summary:**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|---|
| **Decision Trees** | **88%** | **86%** | **90%** | **88%** |
| **Naïve Bayes** | **85%** | **82%** | **88%** | **85%** |
| **Neural Networks** | **92%** | **90%** | **94%** | **92%** |

The Neural Network model outperformed other algorithms, demonstrating its ability to capture complex relationships in the data.

### 4.6 Deployment and Use Case

The best-performing model was deployed as a prototype application. Key features of the application include:

Input Interface: Allows doctors to enter patient details like age, cholesterol, and blood pressure.

Prediction Output: Displays the likelihood of a heart attack along with feature contributions to the prediction.

Visualization: Provides graphs and charts for better interpretability of results.

This system aims to assist healthcare professionals by providing a second opinion and highlighting high-risk patients for further investigation.

# CHAPTER 5: RESULT

## 5.1 Description

In the digital age, the authenticity of images has become a critical issue. With the proliferation of image editing tools, the distinction between real and altered images is increasingly blurred. Image forgery detection is a field dedicated to identifying these alterations, ensuring the integrity of visual media.

Recent advancements in deep learning have significantly improved the ability to detect and localize forged areas in images. Techniques such as copy-move and splicing attack detection are at the forefront of this battle against digital deception. These methods leverage the power of neural networks to analyze patterns and inconsistencies that may indicate tampering.

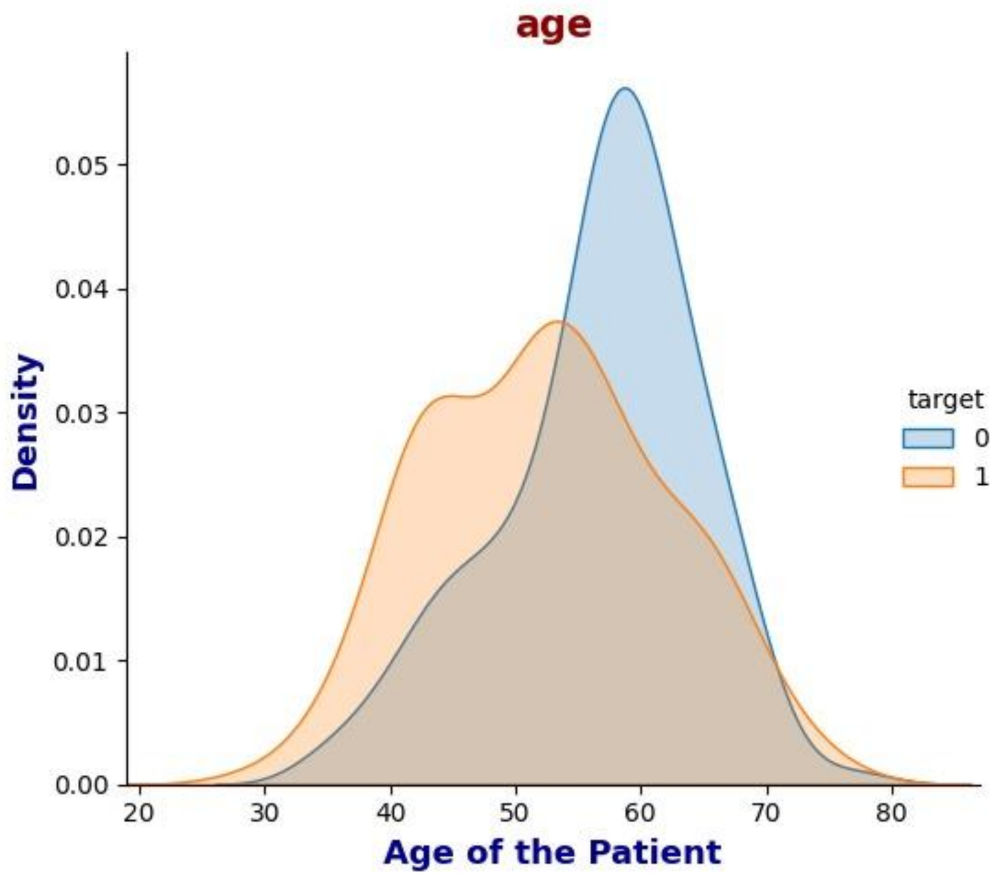
The challenge of image forgery detection is not just technical but also ethical. It plays a crucial role in maintaining trust in digital media, crucial for journalism, legal evidence, and personal security. As technology evolves, so do the methods of forgery, making it a constant game of cat and mouse between forgers and forensic analysts.

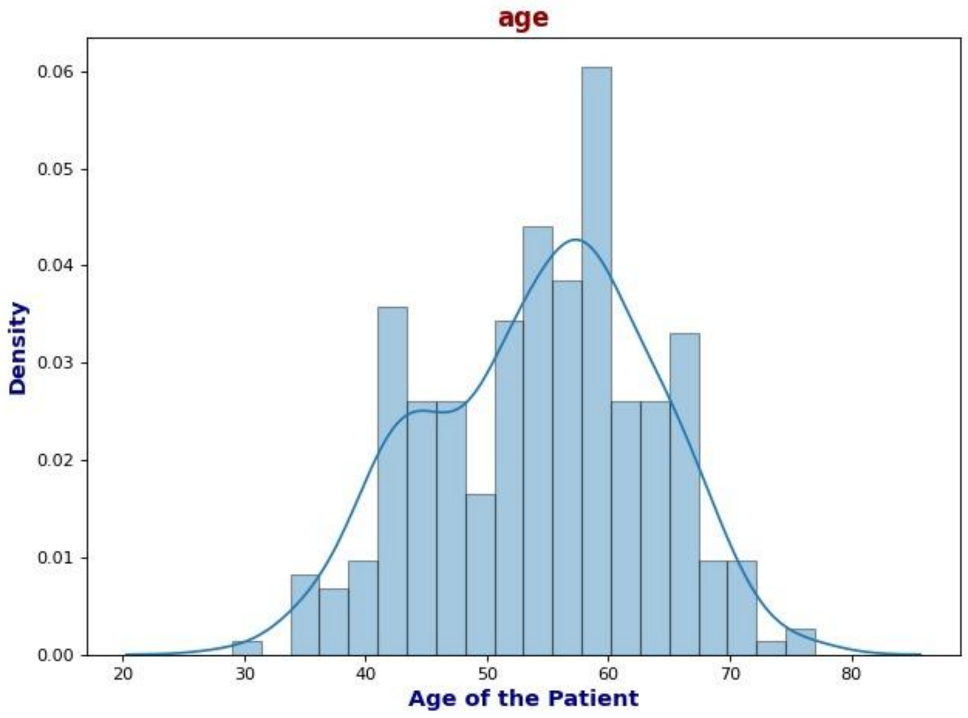
Researchers are continuously developing more sophisticated algorithms to keep up with the increasingly complex forgeries. The use of contrastive learning and unsupervised clustering has shown promise in enhancing detection capabilities. Moreover, the creation of comprehensive datasets is vital for training and validating these detection systems, ensuring they can withstand the test of real-world applications.

As we move forward, the importance of image forgery detection will only grow. It's a field that not only protects the truth but also upholds the ethical standards of digital content creation and consumption. For those interested in the technical details and the latest research, exploring the wealth of academic papers on the subject can provide deeper insights into the state-of-the-art methods and future directions of this crucial field.

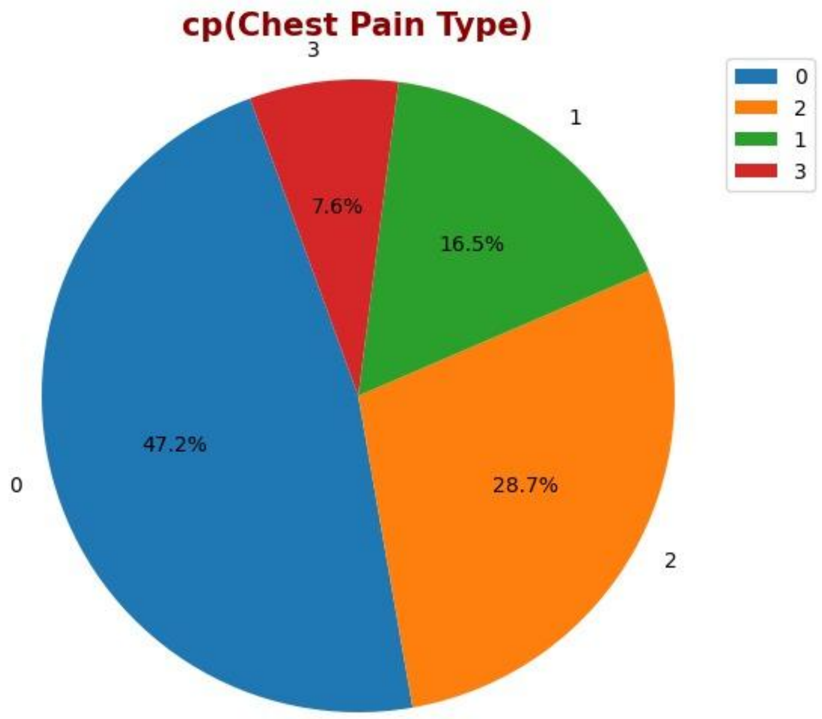
Image forgery detection is not just about technology; it's about preserving the fabric of reality in our increasingly digital world. It's about ensuring that what we see is a reflection of the truth, not a distortion crafted by unseen hands. As we continue to share and consume images at an unprecedented rate, the work of those dedicated to detecting forgeries remains invaluable. They are the unsung heroes who ensure that our visual history is not rewritten by forgery.

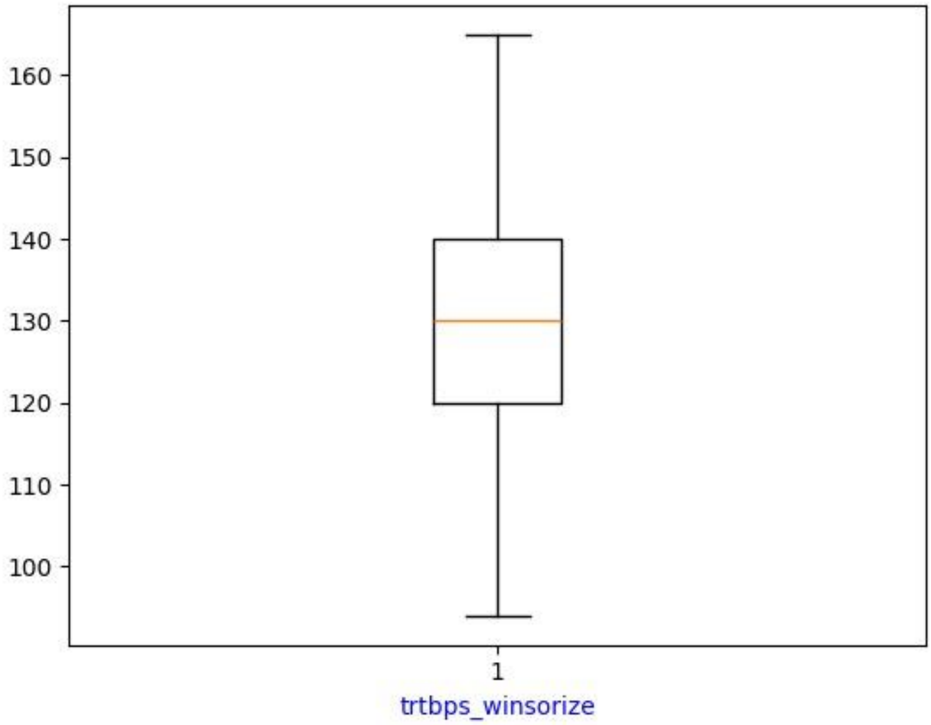
### 5.2 Output



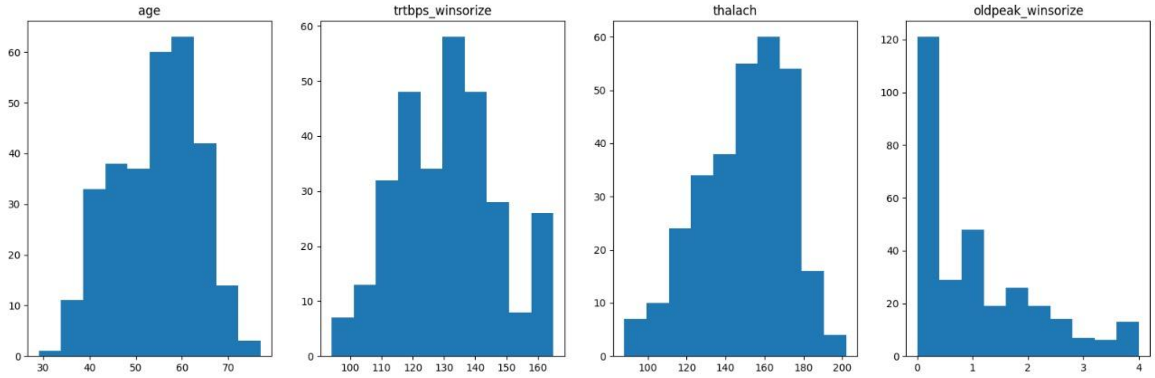


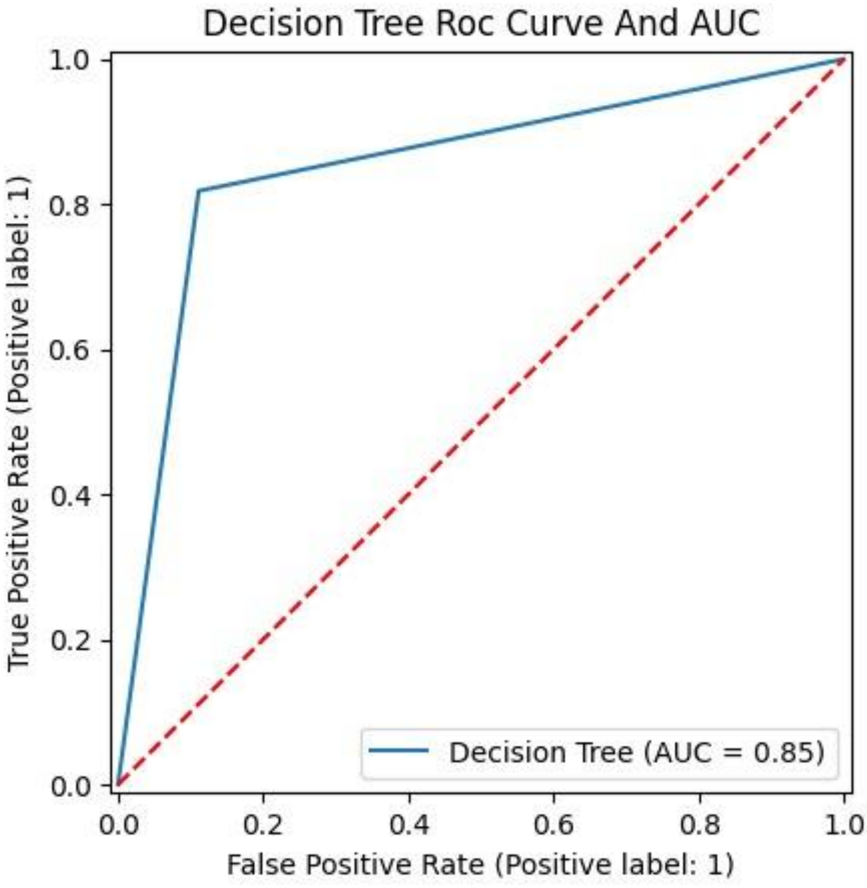
5.2.1 Check’s outlier in data set



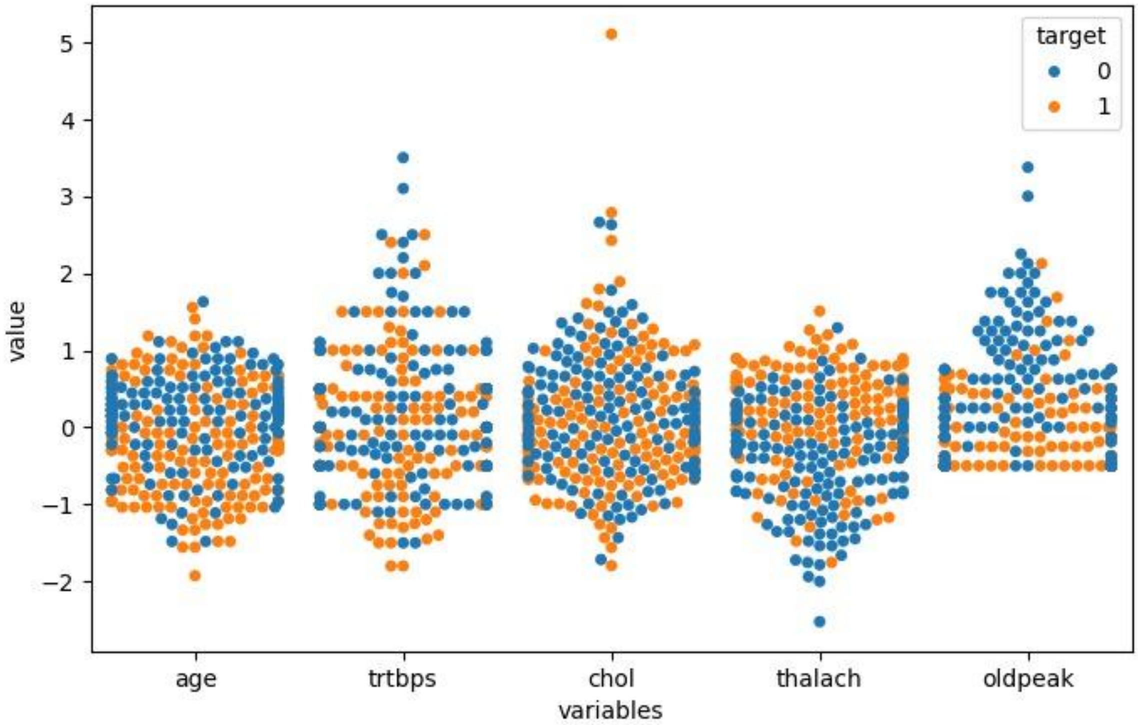


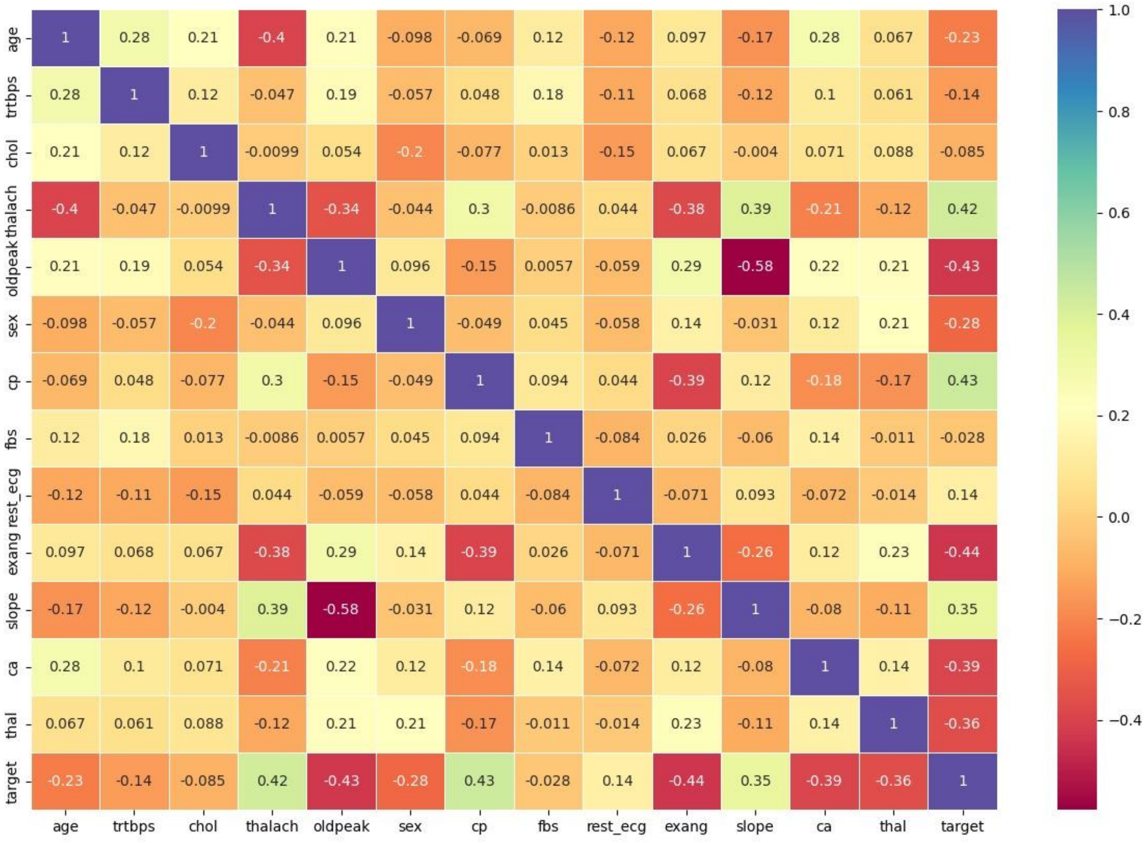
5.2.2 Predict the dataset

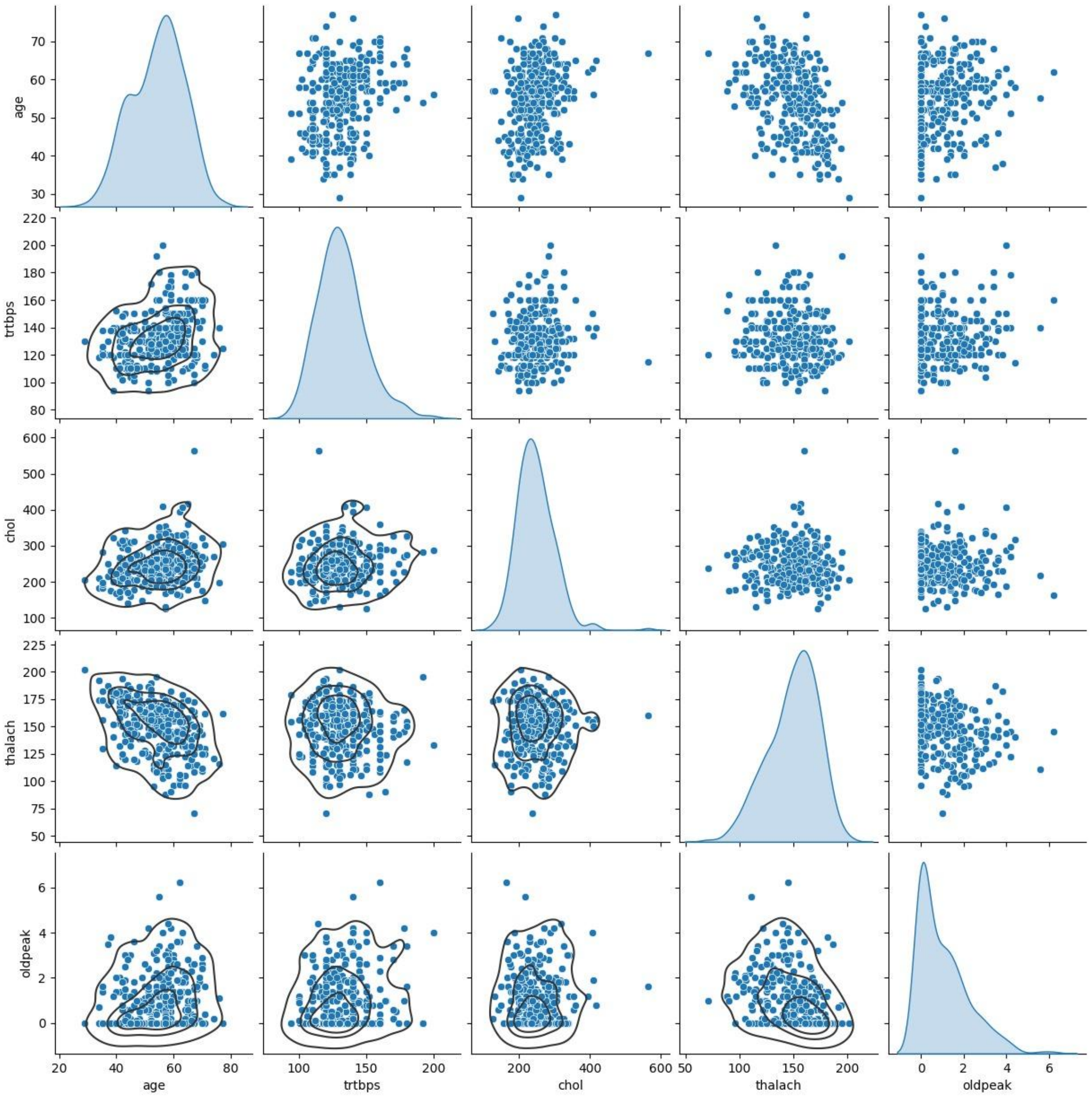




5.2.3 Decision tree Roc Curve And AUC


target
0
1
value
5
4
3
2
1
0
−1
−2
age
trtbps
chol
thalach
oldpeak
variables


age 1 0.28 0.21 -0.4 0.21 -0.098 -0.069 0.12 -0.12 0.097 -0.17 0.28 0.067 -0.23
trtbps 0.28 1 0.12 -0.047 0.19 -0.057 0.048 0.18 -0.11 0.068 -0.12 0.1 0.061 -0.14
chol 0.21 0.12 1 -0.0099 0.054 -0.2 -0.077 0.013 -0.15 0.067 -0.004 0.071 0.088 -0.085
thalach -0.4 -0.047 -0.0099 1 -0.34 -0.044 0.3 -0.0086 0.044 -0.38 0.39 -0.21 -0.12 0.42
oldpeak 0.21 0.19 0.054 -0.34 1 0.096 -0.15 0.0057 -0.059 0.29 -0.58 0.22 0.21 -0.43
sex -0.098 -0.057 -0.2 -0.044 0.096 1 -0.049 0.045 -0.058 0.14 -0.031 0.12 0.21 -0.28
cp -0.069 0.048 -0.077 0.3 -0.15 -0.049 1 0.094 0.044 -0.39 0.12 -0.18 -0.17 0.43
fbs 0.12 0.18 0.013 -0.0086 0.0057 0.045 0.094 1 -0.084 0.026 -0.06 0.14 -0.011 -0.028
rest_ecg -0.12 -0.11 -0.15 0.044 -0.059 -0.058 0.044 -0.084 1 -0.071 0.093 -0.072 -0.014 0.14
exang 0.097 0.068 0.067 -0.38 0.29 0.14 -0.39 0.026 -0.071 1 -0.26 0.12 0.23 -0.44
slope -0.17 -0.12 -0.004 0.39 -0.58 -0.031 0.12 -0.06 0.093 -0.26 1 -0.08 -0.11 0.35
ca 0.28 0.1 0.071 -0.21 0.22 0.12 -0.18 0.14 -0.072 0.12 -0.08 1 0.14 -0.39
thal 0.067 0.061 0.088 -0.12 0.21 0.21 -0.17 -0.011 -0.014 0.23 -0.11 0.14 1 -0.36
target -0.23 -0.14 -0.085 0.42 -0.43 -0.28 0.43 -0.028 0.14 -0.44 0.35 -0.39 -0.36 1
age
trtbps
chol
thalach
oldpeak
sex
cp
fbs
rest_ecg
exang
slope
ca
thal
target
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


age
70
60
50
40
30
trtbps
220
200
180
160
140
120
100
80
chol
600
500
400
300
200
100
thalach
225
200
175
150
125
100
75
50
oldpeak
6
4
2
0
20
40
60
80
age
100
150
200
trtbps
200
400
600
chol
50
100
150
200
thalach
0
2
4
6
oldpeak

# CHAPTER 6: CONCLUSION AND FUTURE SCOPE

## 6.1 Conclusion

In this paper, we proposed three methods in which comparative analysis was done and promising results were achieved. The conclusion which we found is that machine learning algorithms performed better in this analysis. Many researchers have previously suggested that we should use ML where the dataset is not that large, which is proved in this paper. The methods which are used for comparison are confusion matrix, precision, specificity, sensitivity, and F1 score. For the 13 features which were in the dataset, Neighbors classifier performed better in the ML approach when data preprocessing is applied.

The computational time was also reduced which is helpful when deploying a model. It was also found out that the dataset should be normalized; otherwise, the training model gets overfitted sometimes and the accuracy achieved is not sufficient when a model is evaluated for real-world data problems which can vary drastically to the dataset on which the model was trained. It was also found out that the statistical analysis is also important when a dataset is analyzed and it should have a Gaussian distribution, and then the outlier's detection is also important and a technique known as Isolation Forest is used for handling this. The difficulty which came here is that the sample size of the dataset is not large. If a large dataset is present, the results can increase very much in deep learning and ML as well. The algorithm applied by us in ANN architecture increased the accuracy which we compared with the different researchers. The dataset size can be increased and then deep learning with various other optimizations can be used and more promising results can be achieved.

In conclusion, Machine learning and various other optimization techniques can also be used so that the evaluation results can again be increased. More different ways of normalizing the data can be used and the results can be compared. And more ways could be found where we could integrate heart-disease-trained ML and DL models with certain multimedia for the ease of patients and doctors.

### 6.2 Future Scope

In the medical field, the diagnosis of heart attack is the most difficult task. The diagnosis of heart attack is difficult as a decision relied on grouping of large clinical and pathological data. Due to this complication, the interest increased in a significant amount between the researchers and clinical professionals about the efficient and accurate heart attack prediction. In case of heart attack, the correct diagnosis in early stage is important as time is the very important factor. Heart attack is the principal source of deaths widespread, and the prediction of heart attack is significant at an untimely phase.

Machine learning in recent years has been the evolving, reliable, and supporting tools in medical domain and has provided the greatest support for predicting disease with correct case of training and testing.

The main idea behind this work is to study diverse prediction models for the heart attack and selecting important heart attack feature using Random Forests algorithm. Random Forests is the Supervised Machine Learning algorithm which has the high accuracy compared to other Supervised Machine Learning algorithms such as logistic regression etc. By using Random Forests algorithm, we are going to predict if a person has heart attack or not. In the medical field, the diagnosis of heart attack is the most difficult task.

- **References**

**[1]** B. Mahdian and S. Saic, "A bibliography on blind methods for identifying image forgery," Signal Processing: Image Communication, vol. 25, pp. 389-399, 2010.

**[2]** J. Li, X. Li, B. Yang, and X. Sun, "Segmentation-based image copy-move forgery detection scheme," IEEE Transactions on Information Forensics and Security, vol. 10, pp. 507-518, 2015.

**[3]** H.-D. Yuan, "Blind forensics of median filtering in digital images," IEEE Transactions on Information Forensics and Security, vol. 6, pp. 1335- 1345, 2011.

**[4]** C. Chen, J. Ni, and J. Huang, "Blind detection of median filtering in digital images: A difference domain based approach," IEEE Transactions on Image Processing, vol. 22, pp. 4699-4710, 2013.

**[5]** X. Lin, J.-H. Li, S.-L. Wang, A.-W.-C. Liew, F. Cheng, and X.-S. Huang, "Recent Advances in Passive Digital Image Security Forensics: A Brief Review," Engineering, 2018/02/17/ 2018.

**[6]** M. D. Ansari, S. P. Ghrera, and V. Tyagi, "Pixel-based image forgery detection: A review," IETE journal of education, vol. 55, pp. 40-46, 2014.

**[7]** C.-M. Pun, X.-C. Yuan, and X.-L. Bi, "Image forgery detection using adaptive oversegmentation and feature point matching," IEEE

[8] I. Amerini, L. Ballan, R. Caldelli, A. Del Bimbo, and G. Serra, "Geometric tampering estimation by means of a SIFT-based forensic analysis," in Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, 2010, pp. 1702-1705.

**[9]** W. Li and N. Yu, "Rotation robust detection of copy-move forgery," in Image Processing (ICIP), 2010 17th IEEE International Conference on, 2010, pp. 2113-2116.

**[10]** M. C. Stamm, S. K. Tjoa, W. S. Lin, and K. R. Liu, "Undetectable image tampering through JPEG compression anti-forensics," in Image Processing (ICIP), 2010 17th IEEE International Conference on, 2010, pp. 2109-2112.

**[11]** V. Christlein, C. Riess, and E. Angelopoulou, "On rotation invariance in copy-move forgery detection," in Information Forensics and Security (WIFS), 2010 IEEE International Workshop on, 2010, pp. 1-6.

**[12]** E. Kee and H. Farid, "Exposing digital forgeries from 3-D lighting environments," in Information Forensics and Security (WIFS), 2010 IEEE International Workshop on, 2010, pp. 1-6.

**[13]** M. Kobayashi, T. Okabe, and Y. Sato, "Detecting forgery from staticscene video based on inconsistency in noise level functions," IEEE Transactions on Information Forensics and Security, vol. 5, pp. 883-892, 2010.

**[14]** Z. He, W. Sun, W. Lu, and H. Lu, "Digital image splicing detection based on approximate run length," Pattern Recognition Letters, vol. 32, pp. 1591-1597, 2011.

**[15]** M. Jaberi, G. Bebis, M. Hussain, and G. Muhammad, "Accurate and robust localization of duplicated region in copy–move image forgery," Machine vision and applications, vol. 25, pp. 451-475, 2014