# IoT-Enabled Predictive Analytics for Early Detection of CAHD - A Data Driven Framework

**Krishnamurthy H[1] , Ashok Kumar P S[2], Mohan Kumar S[3], Chikka Krishnappa T K[4]**

[1]*Assistant Professor, Dept. of CSE, RVITM, Bengaluru, Karnataka*
[2]*Professor, Dept. od CSE, ACSCE, Bengaluru, Karnataka*
[3]*Professor, Dept. of CSE, MSRIT, Bengaluru, Karnataka*
[4]*Assistant Professor, Dept. of CSE, Sir MVIT, Bengaluru, Karnataka*

## Abstract

The prevalence of CAHD is one of the rise due to lifestyle and genetic factors, for making prepare for the considerable health concern. This paper underwrites to explore the application of various ML approaches to forecast CAHD.

The evaluation is carry out based on cardiac disease datasets, sourced from UCI repositoryon top ofthat data acquired from the IOT-enabled devices. Many performance metrics employed to assess the predictive capabilities where, each ML approaches are recorded in result analysis section, where KNN approach is the best suited to predict CAHD.

The utilization of ML approaches in CAHD prediction can have significant clinical implications. Eventhoug early finding CAHD on individuals will reduces the risk factors and allows timely intervention and preventive measures to patients. Personalized treatment strategies can be designed based on the predicted risk factors, optimizing patient care and improving health outcomes.

**Keywords:** CAHD, ML Algorithms, Performance Metrics, Prediction, IoT

**DOI: 10.48047/ecb/2023.12.8.613**

## Introduction

At Present scenarios, Coronary Artery Heart Disease(CAHD) is one of the fore most reasons for death, as referenced by WHO, mean while, millions of people are ruthlessly infected because of CAHD every year. Heart disease remains a leading cause of mortality worldwide, even though with advancements in medical science and IoT technology, the medical domain has improved a lot even in treatment options in the detection and prevention of diseases in the early stage itself. The advent of IoT has opened up new frontiers in the healthcare domain meanwhile, it revolutionizes how ML approaches are useful while diagnosing and managing cardiovascular conditions and treatment in an early stage.

IoT technology, which connects various devices and sensors through the internet, it is already transformed various techniques to diagnose chronic diseases in the healthcare domain. Even though physicians and consultants are taking advantage of IoT in early predictions of CAHD to save the patients is one of the options. Nowadays, IOT enabled ML approach holds the promise of transforming healthcare by facilitating early prediction of disease and being proactive towards on management of CAHD.

The revolutionary concept of IoT-enabled early prediction of CAHD and its probabilities to be reshaped in the healthcare domain is explored in this paper. By harnessing the power of wearable devices, real-time data analytics, and predictive modeling.So healthcare professionals are expert enough to classify the individual's risk, initiation, and interventions towards symptoms and chronic manifestation.

CAHD is one of the most trivial types of disease, it can affectthe heart very badly and it causes heart attack or heart arrest immediately, even though timely treatment and first aid can save the CAHD patients. Many features will affect the structure and functioning of heart, it cause CAHD, so physicians and consultants will face huge confusion to find the heartdisease accurately and rapidly. The significance of the CAHD predictive model is to find the CAHD in early state itself with very low cost. As per the survey, many people are facing unexpected death without any symptoms,even though; people are silently sufferingfrom CAHD.

IoT (Internet of Things) is an advanced technology, it has incredible potential in various fields, including healthcare domain. Using IoT technology, early prediction of heart disease can be highly benefitted to improve patient outcomes and reducing healthcare costs, like, Wearable Devices, Remote Patient Monitoring, Data Analytics and Machine Learning, Predictive Models, Alerts and Notifications, Integration with EHR, Patient Engagement.

Generally across the world, health workers may found many risk factors, it is very highly responsible and escalating the threat towards on CAHD. Various risk factors and human habitual things like alcoholism, smoking, tobacco, highblood pressure, stress, cholesterol, diabetes, obesity, overweight, creatinine level,*etc*. it will boost the possibility of CAHD, even it is potential to control, manage and modify the risk factors, So,the person’s might have a improved possibility to save his life by CAHD.

Ultimately, the goal of IoT-enabled early prediction of CAHD, is to shift the focus of healthcare from reactive treatment of heart conditions to proactive prevention and early intervention. By leveraging the power of IoT technology and predictive analytics, healthcare providers can significantly improve cardiovascular health outcomes and enhance the overall quality of life for individuals at risk of heart disease.

Many researchers are studied and revealed that machine learning techniques have better thoughtfulness, conclusion, andaccuracy while examining patient classification problems based on the data model. Various classification algorithms are to be utilized in the CAHD classification problems, where various algorithms are SVM, RF, NB, and KNN are utilized in the early CAHD prediction.

## Literature Review

The primary objective of IoT-enabled early prediction of heart disease is to improve patient outcomes and reduce the burden of heart-related health issues by identifying potential cardiovascular risks at an early stage.

Problem statement:

- Early prediction of CAHD using Machine Learning algorithm.
- Develop an various approaches to afford QoS in health care domain using IoT devices.

In general, the classification of data is one of the most decisive things, for example, predicting coronary heart disease accurately and efficiently is a big challenge, even though some of the machine learning techniques have been used and described as follows;

Analysis and Prediction of diseases are two most demanding factors to be faced critically by the doctors and data scientist, where data analytics be very delightful issue, so in this regard, many health industries will working on variety of human syndromes, where they generate huge data. Heart disease, cancer, tumour and Alzheimer's disease are one of the chronic human diseases, where data scientist and doctors are doing rapid and efficient analysis on these diseases using many machine learning techniques to study and predict these diseases to save and reduce human deaths [1].

Heart disease is a significant medical issue and it influences countless individuals. Cardiovascular Disease (CVD) is one such danger. CVD is a significant health issue that affects entire world, particularly in low as well as middle income countries. To extract valuable data in case of heart disease prediction, different types of machine learning as well as data mining methods are executed. However, the accuracy of the ideal results is not acceptable. Using machine learning techniques, this paper suggests risk prediction system of CVD. his paper discusses briefly about systems major modules along with related theory. The proposed model consolidates AI and information mining to give the exact results with least mistakes [2].

proposed for heart disease prediction feature extraction approaches and classification using ensemble deep learning. First, Feature extraction using SIFT and ALEXANET from the Mask Region-Based Convolutional Neural Network **(**RCNN) instance segmented image. Second one, Hybrid Classification with the combination of Random forest and Gaussian Navies Bayes to detect the heart attack [3].

Paper proposed to introduce necessary data analysis techniques likes decision tree technique using KEEL tool, and finally, we analyze and classify the structure and graphical pattern of the data, with the help of machine learning methodology and graph theory. Eventually, our tailored method is finalized with decision tree for validation purpose. The simulation results of our approach on different databases show the feasibility and effectiveness of our proposed framework [4].

Machine learning to categorize cardiac illness based on electrocardiogram (ECG) and phonocardiogram (PCG) readings. The investigation began with signal preprocessing, which included cutting and normalizing the signal, and was accompanied by a continuous wavelet transformation utilizing a mother wavelet analytic morlet. The results of the decomposition are shown using a scalogram, and the outcomes are predicted using the hidden semi morkov model

(HSMM). A modal is built for the one with the max frequency by training the dataset to it. In first step, input for each cardiac parameter is provided, and the sick stage of the heart is predicted based on the modal created. We then take measures based on the patient's condition [5].

While characterizing with forecasting and analysis of various heart-related syndromes in patients with wide range of age by means of machine learning algorithms and techniques. In this case study, many parameters are considered to do analysis and predict heart disease of patients, where KNN, logistic regression and decision tree algorithm are used to calculate accuracy and performance [6].

In medical domain gather information from past experiments and determine them to find trends and results to present situations. For the exact prediction, heterogeneous ensemble classifier is proposed for the early stage diagnosis. It identifies causes, symptoms and estimates the treatment for the Dilated Cardiomyopathy, which is due to the heart muscle inflection. The proposed system will be significant with prognosis analysis for the exact prediction and diagnosis of the dilated cardiomyopathy and heart failure. Improved Genetic algorithm (IGA) model generates greater accuracy level for all HD compared to other classifiers [7].

To developing a Web-based Heart Disease Prediction System (HDPS) byapplying DT and NB ML algorithms. We are using the UCI repository HD dataset to traina model by comparing DT and NB algorithm for HDPS Web application. The datasetcontains 303 instances with 14 attributes that help to train a prediction model that will bedeployed into a web application for prediction. The main aim of this project is to build anefficient prediction model and deploy for prediction of disease. An HDP Model is built byusing NB algorithm that provides 88.163% accuracy among others. A web-based HDPSapplication is developed through the waterfall model. Each phase is efficiently done [8]

## 1. System Architecture

To implement an IoT-enabled early prediction CAHD system requires a well-structured methodology that incorporates data collection, analysis, predictive modeling, and intervention strategies.

**Objective of the research is:**

- ❖ Design and develop a novel algorithms used to handle classificationand prediction of CAHD in early stage.

- ❖ To estimate and predict abnormalities in the health conditionof patients using wearable sesor devices.

Following phases are required in the proposed methodology:

1. Data acquisition,
2. Data preprocessing,
3. Implementation of the techniques,
4. Performance evaluation.

**Dataacquisition phase**

Heart disease dataset tis acquired from real time medical sensors and from the UCI repository database also, both male and female persons data to be considered in the dataset. Some and around 270 samples are considered, where every sample is splitted into13 attributes, the details of the data set is shown in Table1.
In the sample dataset can be segregate into,

- Negative class - A samplewith absence of CAHD
- Positive class - A samplewith presence of CAHD.

Data correlation is the way in which one set of data may correspondto another set. In a machine learning perspective, howthefeaturescorrespondtooutput.Sometimes,itbecomesvery hard to figure out how the data are correlated witheach other. In this situation, data visualization can help to find out how individual features may correlate with the output. Pearson's Correlation Coefficient helps to find out the relationship between the two quantities. The value of Pearson's Correlation Coefficient can be between −1to +1, where 1 means that they are highly correlated and 0means no correlation, −1 means that there is a negativecorrelation.

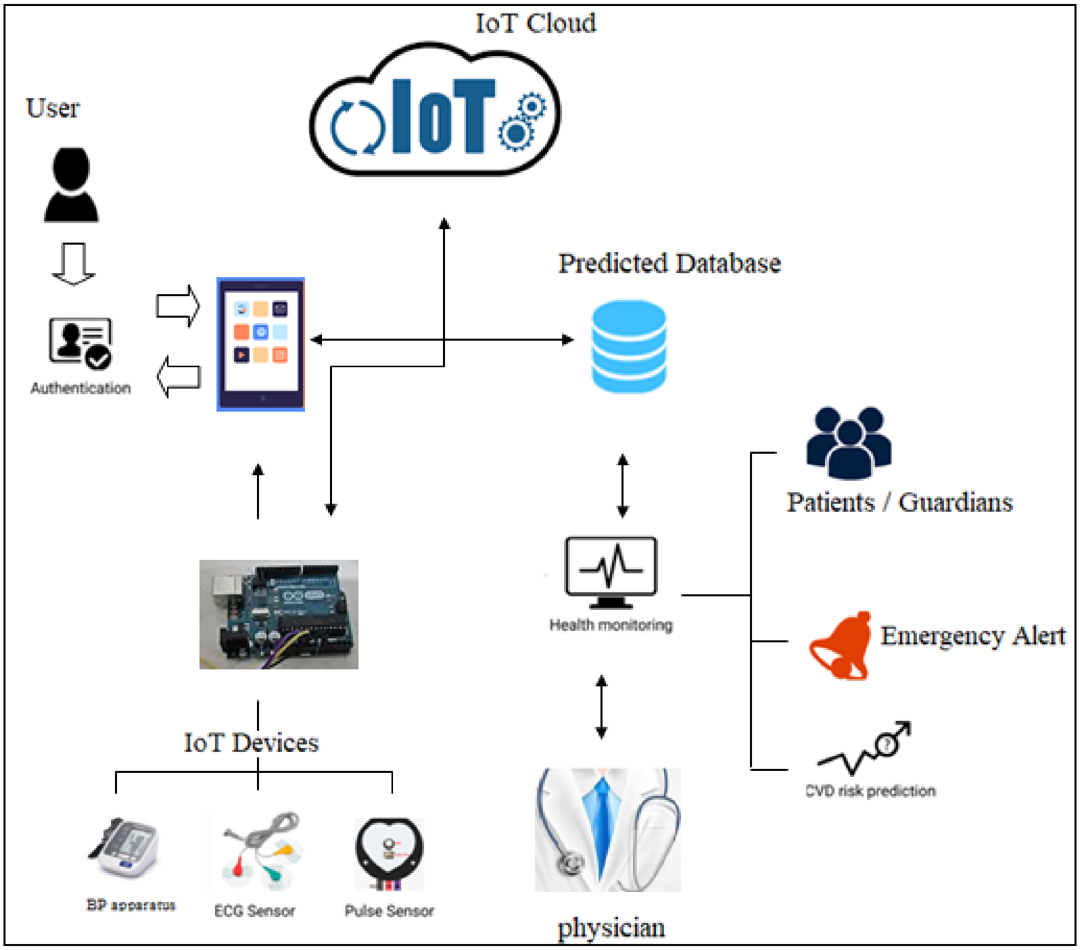


Fig 1- Architectural diagram of the Framework

An Android mobile application has been developed for end users, and an Arduino module is used to collect data from end users using various medical sensors, which is then stored in the cloud. The best performing ML model from the Predictive Analysis Module is used in the health monitoring module. The application was created in Android Studio based on a prototype created in Figma, a collaborative browser-based interface design tool. The user interface module allows users to upload their profile and real-time data to the Predicted database. Based on the data collected from the user, the best performing ML model, deployed in the IBM cloud, then formulates the prediction result.

The interaction module also allows users to check their health status in real time and, if necessary, initiate emergency alerts. Consultations gathered from doctors in this field are also made available to users based on their CAHD risk level. Figure 1, depicts the proposed system's System Architecture.

Users are associated with the application's features after a proper authentication process. Upon the first login, some profile data such as weight, cholesterol, fasting blood sugar, and type of angina during exercise (if any) is collected from users, with the option to update them as needed. This application allows the user to control hardware components via Bluetooth technology, allowing for real-time data collection. Users must connect all three devices sequentially, and readings from the Data Acquisition Module will be collected and saved in the Firebase data repository. Users can see their risk of developing cardio-vascular diseases based on this information.Based on their preferences, users can classify the prediction result into two categories.

For risk level classification, results are divided into green, yellow, and red zones, indicating a low to high risk of CAHD. In the case of two zone classification, users are divided into red and green zones based solely on the presence or absence of CAHD.

This framework also allows users to track their health in real time. This application can monitor your heart condition using ECG signals. There are also options for monitoring blood pressure and pulse, and in the event of an emergency, appropriate alerts will be issued. Users will also be provided with the necessary User interface for the device Connection based on their health condition.

- **Data Set:**

There are several features or factors that have been found to be correlated with cardiovascular disease (CAHD). These features can help identify individuals who may be at a higher risk of developing CAHD.

Table 1 – descriptions of the Features in dataset

| Feature | Attributename | Domain |
|---|---|---|
| Age | age | Ageinyears:29-77 |
| Sex | sex | Male=1,Female=0 |
| Chestpaintype | cp | 1=typicalangina;<br>2=atypical angina;<br>3=non-anginalpain |

| | | |
|---|---|---|
| RestingBloodPressureinmm/Hg | trestbps | 94-200 |
| SerumCholesterolinmg/dl | chol | 126-564 |
| Fastingbloodsugar>120mg/dl | fbs | 1=yes,0=no |
| RestingECGobservations | restecg | 0=normal; |
| Maximum heartrateachieved | thalach | 71-202 |
| STdepressioninducedbyanginarelative | oldpeake | 0-6.2 |
| Exercise induced angina | exang | 1=yes,0=no |
| Slope of the peak exercise ST segment | slope | 1=up sloping<br>2= flat |
| Thallium stress test result | thal | 3 = normal; 6 = fixed defect; |

Table 1 represents the description of the 14 different features required in dataset.

Some of the most significant features include [1]:

- Age: The risk of CAHD increases with age. As individuals get older, the likelihood of developing heart disease, high blood pressure, and other cardiovascular problems increases.
- Gender: Men are generally at a higher risk of developing CAHD compared to premenopausal women. However, after menopause, women's risk of CAHD catches up to that of men.
- Family history: Having a close family member, such as a parent or sibling, with a history of CAHD increases an individual's risk. This suggests a potential genetic predisposition to developing heart disease.
- Ethnicity: Certain ethnic groups, such as African Americans, Hispanics, and South Asians, have a higher prevalence of CAHD compared to others. This may be due to a combination of genetic, lifestyle, and socioeconomic factors.
- High blood pressure (hypertension): Hypertension is a significant risk factor for CAHD. Individuals with high blood pressure have an increased risk of heart attack, stroke, and other cardiovascular problems.
- High cholesterol levels: Elevated levels of LDL cholesterol (bad cholesterol) and low levels of HDL cholesterol (good cholesterol) are associated with an increased risk of CAHD. High cholesterol can lead to the formation of plaque in the arteries, narrowing them and reducing blood flow to the heart.
- Obesity: Being overweight or obese is linked to a higher risk of developing CAHD. Excess body weight can contribute to conditions such as high blood pressure, high cholesterol, and diabetes, all of which increase the chances of heart disease.

- ❖ Diabetes: People with diabetes are more likely to develop CAHD. High blood sugar levels can damage blood vessels and nerves, leading to a higher risk of heart attack, stroke, and other cardiovascular complications.
- ❖ Smoking: Smoking tobacco is a significant risk factor for CAHD. It damages blood vessels, reduces oxygen supply to the heart, increases blood pressure, and promotes the formation of blood clots, all of which contribute to the development of heart disease.
- ❖ Sedentary lifestyle: Lack of physical activity is associated with an increased risk of CAHD. Regular exercise helps maintain a healthy weight, lowers blood pressure, improves cholesterol levels, and reduces the risk of developing diabetes, all of which are beneficial for heart health.

It's important to note that these features are correlated with CAHD, but correlation does not necessarily imply causation. However, identifying and addressing these risk factors can help reduce the risk of developing cardiovascular disease.

- **Correlations about the features:**

The correlation of features for cardiovascular disease (CAHD), we are looking at how different factors or characteristics are related to the presence or risk of developing CAHD.
For example, if we say there is a positive correlation between high blood pressure and CAHD, it means that as blood pressure levels increase, the likelihood of developing CAHD also increases. On the other hand, if there is a negative correlation between physical activity and CAHD, it means that as physical activity levels increase, the risk of CAHD decreases.

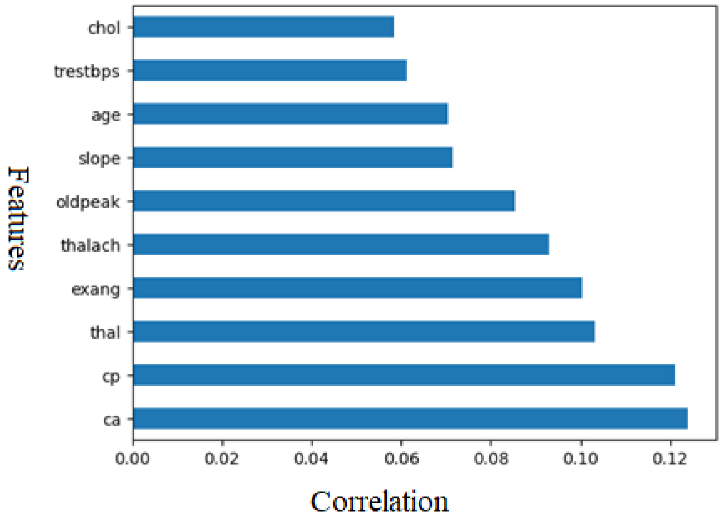


Fig 2 – Correlation between Heart disease with features

Correlation can be measured using statistical techniques such as correlation coefficients, which range from -1 to +1. A positive correlation coefficient indicates a positive relationship, where both variables increase or decrease together. A negative correlation coefficient indicates an inverse relationship, where one variable increases while the other decreases. A correlation coefficient of 0 suggests no linear relationship between the variables.

Understanding the correlation of features for CAHD can help in identifying risk factors and developing preventive strategies. However, it's important to note that correlation does not imply causation. While two variables may be correlated, it does not necessarily mean that one directly causes the other.

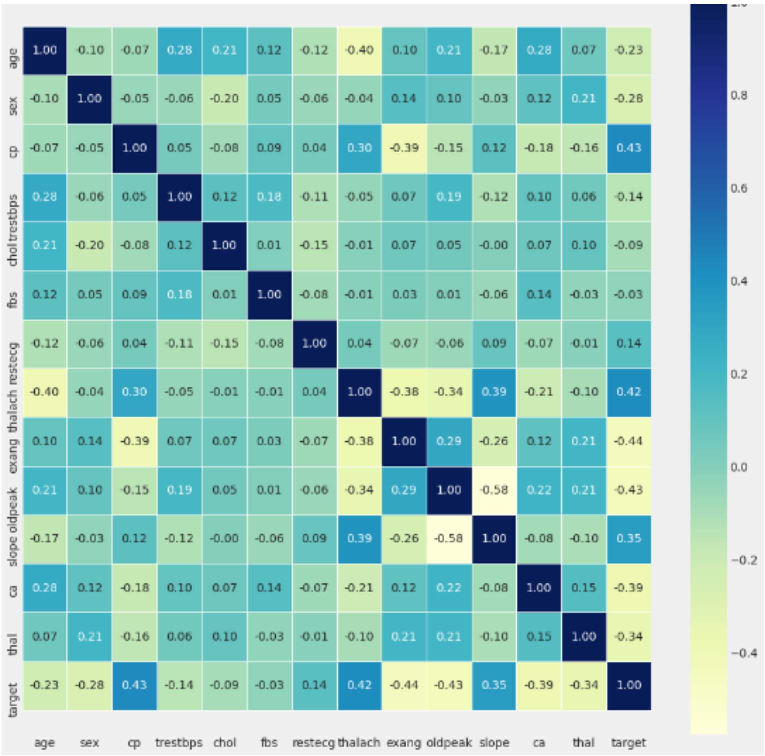


Fig 3 – Correlation matrix with features

- **DataPreprocessing**

  Data preprocessing is an essential step in the data analysis channel, it involves transforming raw data into a format that is suitable for analysis and machine learning models. The goal of data preprocessing is to clean, normalize, and transform the data to improve the quality and effectiveness of subsequent analysis or model training, shown in figure 4.

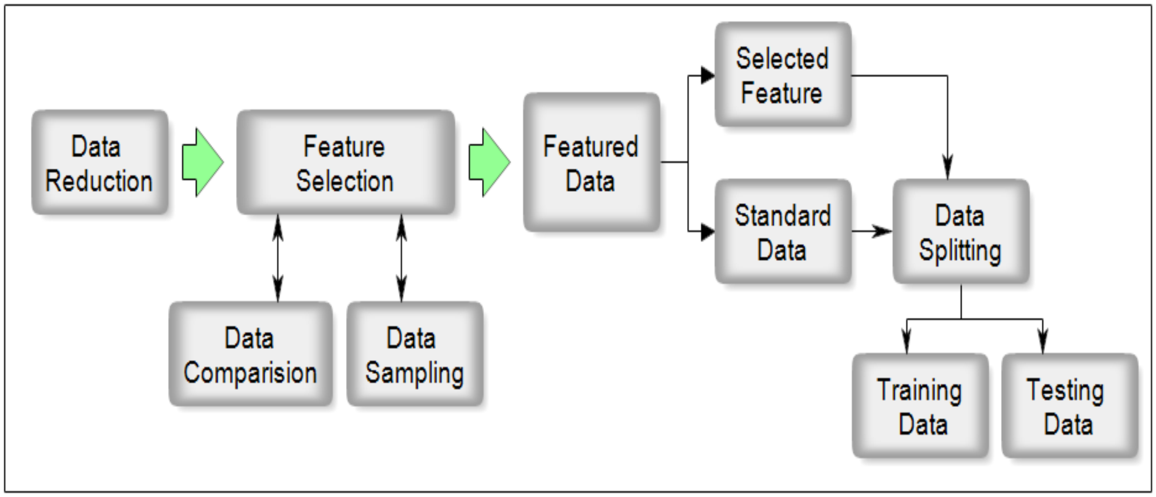


Fig 4 – Data Preprocessing stages

Here are some common techniques used in data preprocessing [7]:

- Data Cleaning: This involves handling missing values, outliers, and inconsistent data. Missing values can be imputed or removed based on the analysis. Outliers can be detected and treated by either removing them or applying statistical techniques. Inconsistent data can be resolved by standardizing formats or correcting errors.

- Data Transformation: This step involves transforming the data to meet certain assumptions of the analysis or model. Common transformations include scaling numerical features, encoding categorical variables, and handling skewed distributions through techniques like logarithmic or power transformations.

- Feature Selection: It involves selecting the most relevant features for the analysis or model. This can be done using statistical techniques, domain knowledge, or feature importance measures from machine learning models.

- Dimensionality Reduction: When dealing with high-dimensional data, dimensionality reduction techniques like Principal Component Analysis (PCA) can be applied to reduce the number of features while preserving important information.

- Data Normalization: Scaling numerical features to a standard range between 0 and 1 can help avoid bias in models that are sensitive to the scale of the data. Common normalization techniques include min-max scaling and z-score normalization.

- Splitting Data: It is important to split the dataset into training, validation, and testing sets. The training set is used to train the model, the validation set is used for hyperparameter tuning, and the testing set is used to evaluate the final model's performance.

- **Machine Learning Techniques**

Machine Learning technique is a collection of computational models. Using it many kinds of problems can be solved easily which are difficult to solve usingconventional computational algorithms [8].

Employ machine learning algorithms, such as logistic regression, support vector machines, random forests, or deep learning techniques, to build predictive models. Train these models using historical patient data, incorporating the engineered features.

There are various applications of computational intelligence in the fields of medicine and pathology. The SVM, DT, NB, RF, and KNN algorithms used in this comparative 4 study are explained in detail in this section. Reinforcement learning is another intriguing method [10].

For classification problems with imbalanced datasets, Reinforcement Learning performs better. For Reinforcement learning has been used to remove noisy data and learn classification problems. better features, which resulted in a significant improvement in classification performance. However, this The technique works best with large amounts of data and images [11].

- **SupportVectorMachine**

The basic idea behind SVM is to find the optimal hyperplane that best separates data points of different classes in the feature space.

For the classification problem, given a set of labeled data points $(x_i, y_i)$, where $x_i$ represents the feature vector of the i-th data point, and $y_i$ is the corresponding class label (either -1 or +1 for binary classification), the SVM aims to find a hyperplane in the feature space defined as:

$(w \cdot x + b) = 0$ --- (1)

where w is the weight vector perpendicular to the hyperplane, and b is the bias term. The parameter w and b define the decision boundary that separates the data into different classes. The goal is to maximize the margin between the two classes, which is the distance between the closest data points (support vectors) to the decision boundary.

The optimization problem for SVM can be formulated as:

minimize: $1/2 \, ||w||^2 + C \, \Sigma \, \xi_i$ ----- (2)

subject to: $y_i(w \cdot x_i + b) \geq 1 - \xi_i$ ------ (3)

for all data points $(x_i, y_i)$

where $||w||$ is the L2 norm of the weight vector w,

C is the regularization parameter that controls the trade-off between maximizing the margin and minimizing the classification error, and $\xi_i$ are slack variables that allow for some misclassification (soft margin).

For non-linearly separable datasets, SVM employs a kernel trick to map the data into a higher-dimensional feature space where the data points become separable. Commonly used kernel functions include polynomial kernel, radial basis function (RBF) kernel, and sigmoid kernel.

To make predictions on new data points, the class label is determined based on the sign of the decision function:

$f(x) = (w \cdot x + b)$ ---- (4)

If $f(x) \geq 0$, the data point x belongs to class +1.
If $f(x) < 0$, the data point x belongs to class -1.

SVM's mathematical model allows it to find the optimal hyperplane, even in cases where the data points are not linearly separable. Its ability to handle complex data distributions and high-dimensional feature spaces makes SVM a versatile and widely used algorithm in various applications, including image recognition, text classification, and bioinformatics.

- **Logistic Regression**

Logistic Regression is a widely used statistical and machine learning algorithm used for binary classification problems, where the target variable has two classes (e.g., 0 and 1, True and False, Yes and No). The mathematical model of logistic regression is based on the logistic function (also known as the sigmoid function), which maps any real-valued number to the range [0, 1].

For binary classification, we have a dataset with labeled data points $(x_i, y_i)$, where $x_i$ represents the feature vector of the i-th data point, and $y_i$ is the corresponding binary class label (0 or 1). The logistic regression model aims to find the optimal set of parameters (weights and bias) that best separates the two classes.

The logistic regression model is defined as follows:

Hypothesis Function: The hypothesis function h(x) is the core of logistic regression and uses the logistic (sigmoid) function to transform the linear combination of features (x) and their corresponding weights (θ) into a probability value:

$h(x) = 1 / (1 + \exp(-\theta \cdot x))$ ----- (5)

Here, exp() is the exponential function, and θ is the parameter vector (weights) that we want to learn.

- **DecisionTree**

For binary classification, given a set of labeled data points $(x_i, y_i)$, where $x_i$ represents the feature vector of the i-th data point, and $y_i$ is the corresponding binary class label (0 or 1), the decision tree algorithm works as follows:

Selecting the Best Feature: The algorithm selects the best feature from the available features to split the data at the root node. It chooses the feature that provides the most significant reduction in impurity or increase in information gain, which is typically measured by metrics like Gini impurity or entropy.

Splitting the Data: The data is split into subsets based on the selected feature. Each subset corresponds to one of the possible outcomes of the feature. This process is repeated recursively for each subset until a stopping criterion is met (e.g., reaching a maximum depth, minimum number of samples per leaf, or impurity threshold).

Assigning Class Labels to Leaf Nodes: At each leaf node, the majority class label of the data points in that leaf is assigned as the predicted class label for that region of the feature space.

Handling Continuous Features: For continuous features, decision tree algorithms perform binary splits (e.g., $x >$ threshold, $x <=$ threshold) to divide the data into subsets.

- **NaïveBayes**

The Naive Bayes classifier is a simple probabilistic model based on Bayes' theorem with an assumption of independence between features. The mathematical model of Naive Bayes involves calculating the posterior probability of each class given the feature vector using the Bayes' theorem.
Given a set of training data points (x_i, y_i),
where x_i is the feature vector and y_i is the corresponding class label.

the Naive Bayes model can be represented as:
P(y|x) = P(x|y) * P(y) / P(x) ----- (6)

where P(y|x) is the posterior probability of class y given the feature vector x,

P(x|y) is the likelihood of the feature vector x given class y,
P(y) is the prior probability of class y, and
P(x) is the evidence probability.

The Naive Bayes assumption assumes that all features in the feature vector are conditionally independent given the class label.

This assumption allows us to factorize the likelihood term as:
P(x|y) = P(x_1|y) * P(x_2|y) * ... * P(x_n|y) ---- (7)

where P(x_i|y) is the probability of feature x_i given class y.

## Experimental Details

- **Performance metrics**

Performance metrics are used to evaluate the effectiveness and accuracy of machine learning models. They provide quantitative measures of how well a model is performing in terms of various aspects such as classification accuracy, error rates, precision, recall, and F1 score.

Using the confusion matrix, we can visualize the performance of ML approaches. In the confusion matrix, four classification performance are represented.

*TP* = True Positive (Correctly Identified)
*TN* = True Negative (Incorrectly Identified)
*FP*=False Positive (Correctly Rejected)
*FN*=False Negative(Incorrectly Rejected)

To evaluate the performance of the system, the following qualities are measured.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad \text{----- (8)}$$

$$Sensitivity = \frac{TP}{(TP+FN)}$$ -------------- (9)

$$Specificity = \frac{TN}{(TN+FP)}$$ ----------------- (10)

$$Precision = \frac{TP}{TP+FP}$$ -------------- (11)

$$Negative\ Predictive\ Value(NPV) = \frac{TN}{FN+TN}$$ ----------- (12)

$$F1\ Score = \frac{2TP}{(2TP+FP+FN)}$$ ---------------- (13)

In this paper, many ML approaches like, SVM, DT, NB, RF, and KNN are used. The CAHD database, contain 270 samples, while diagnosing the data can be classified into two types, one is healthy persons and other type is a patient, who is to be a possible CAHD victim, i.e., 150 samples could be indicated that the absence of heart disease and 120 samples with the incidence of heart disease.

**ROC**

In cardiovascular disease (CAHD), ROC analysis can be used to evaluate the performance of a classification model in predicting the presence or absence of CAHD based on certain risk factors or biomarkers.

The ROC curve plots the true positive rate (sensitivity) against the false positive rate (specificity) at various threshold settings of the classification model. The area under the ROC curve (AUC) is a commonly used metric to quantify the overall performance of the model. A higher AUC indicates a better discriminatory power of the model in distinguishing between individuals with and without CAHD.

By analyzing the ROC curve and calculating the AUC, researchers and clinicians can assess the accuracy, sensitivity, specificity, and overall predictive power of a classification model for CAHD. This information can be valuable in evaluating the effectiveness of risk assessment models, diagnostic tests, or predictive algorithms in the field of cardiovascular medicine.

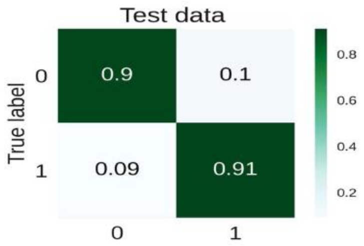


Fig 5- ConfusionMatrix for test dataPerformance

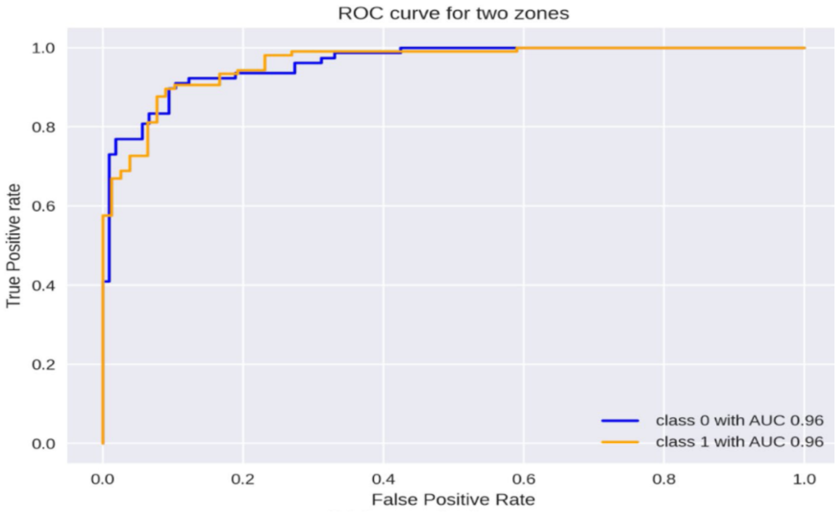
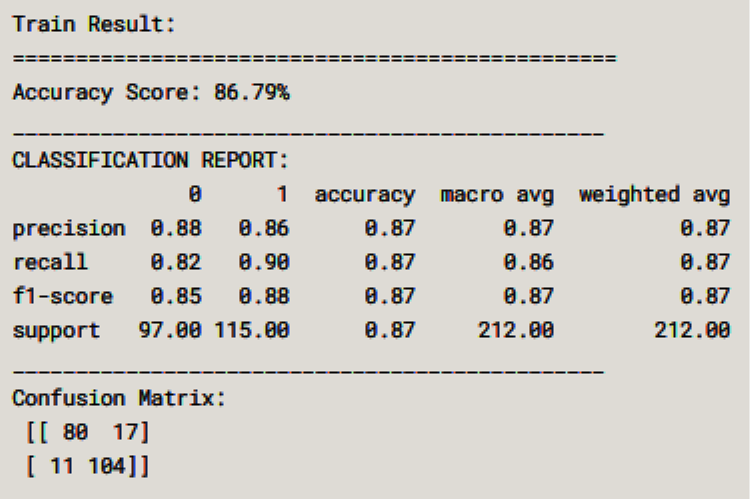


Fig 6 – performance of ROC-AUC for two level classification

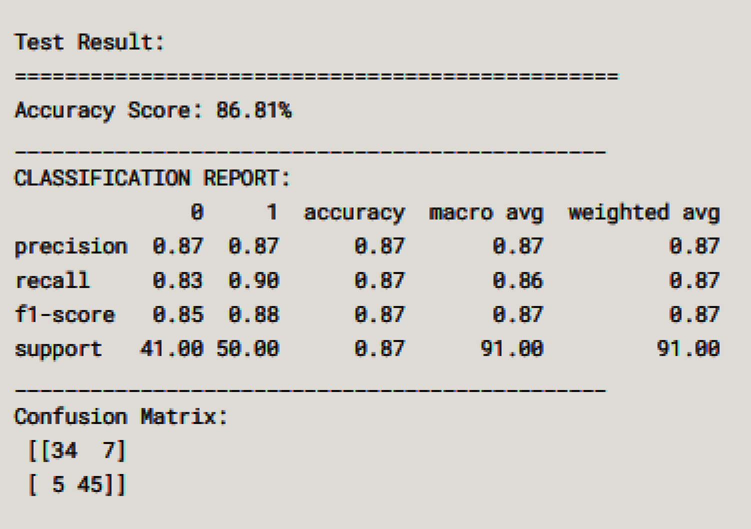
## 4. Result Analysis

Three approaches were used: applying different machine learning algorithms and then using deep learning to see what difference it makes when applied to data. In the first approach, the acquired normal dataset is directly used for classification, whereas in the second approach, the data with featureselection is handled and no outliers are detected. The results obtained are quite promising, and in the third approach, the dataset was normalised while taking into account outliers and feature selection; the results obtained are much better than the previous techniques, and when compared to other research accuracies, our results are quite promising.



```
Train Result:
================================================
Accuracy Score: 86.79%
_______________________________________________
CLASSIFICATION REPORT:
              0      1  accuracy  macro avg  weighted avg
precision  0.88   0.86      0.87       0.87          0.87
recall     0.82   0.90      0.87       0.86          0.87
f1-score   0.85   0.88      0.87       0.87          0.87
support   97.00 115.00      0.87     212.00        212.00
_______________________________________________
Confusion Matrix:
 [[ 80  17]
 [ 11 104]]
```

Table 2 – Training Result of the model



```
Test Result:
================================================
Accuracy Score: 86.81%
_______________________________________________
CLASSIFICATION REPORT:
              0     1  accuracy  macro avg  weighted avg
precision  0.87  0.87      0.87       0.87          0.87
recall     0.83  0.90      0.87       0.86          0.87
f1-score   0.85  0.88      0.87       0.87          0.87
support   41.00 50.00      0.87      91.00         91.00
_______________________________________________
Confusion Matrix:
 [[34  7]
 [ 5 45]]
```

Table 3 – Test result of the model

These performance metrics help researchers and clinicians to assess the accuracy, sensitivity, specificity, precision, and overall predictive power of models or algorithms in predicting the presence or absence of CAHD. The proposed system uses various algorithms like, SVM, NB,RF, DT, KNN. Based on the data set, above mentioned ML approaches are used and gain performance metrics like, Precision, Recall and F1-Score, where KNN approach has good performance compre to other approach. Figure 7, 8,9 shows the performance analysis of all other factors like, Precision, Recall and F1-Score. While evaluating the performance metrics based on the various models, KNN approach is the best in metrics to predict CAHD in early stage .

Table 4 - Evaluation of models with performance metrics are,

| Models | **Precision** | **Recall** | **F1score** |
|---|---|---|---|
| KNN | 0.877 | 0.821 | 0.848 |
| NaiveBayes | 0.739 | 0.705 | 0.784 |
| RandomForest | 0.781 | 0.78 | 0.819 |
| SupportVectorMachine | 0.805 | 0.745 | 0.785 |
| DecisionTree | 0.708 | 0.78 | 0.829 |

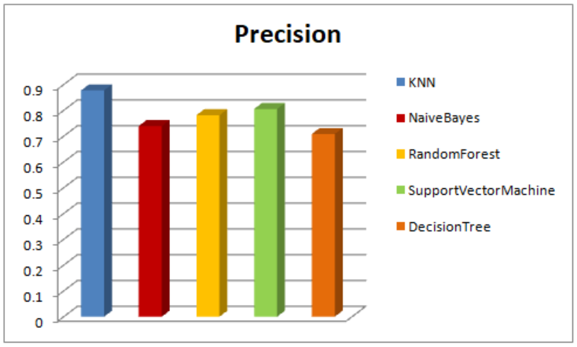


Fig 7 – Performance analysis on Precision

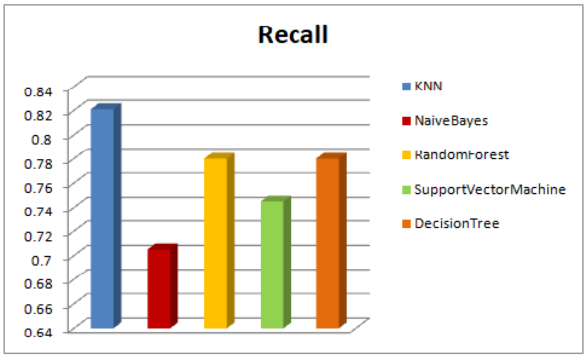


Fig 8 – Performance analysis on Recall

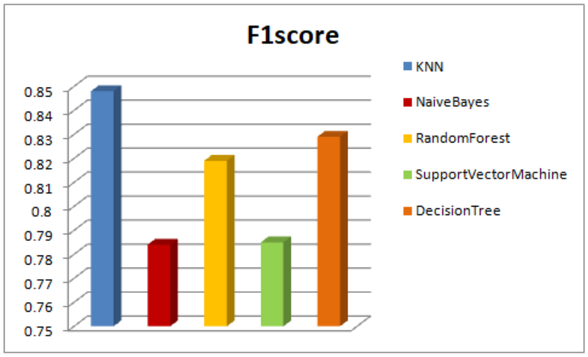


Fig 9 – Performance analysis on F1 score

## CONCLUSION

IoT-enabled early prediction of heart disease represents a transformative paradigm shift in cardiovascular healthcare. By leveraging the power of IoT technology, wearable devices, data analytics, and predictive modeling, this approach holds immense potential in revolutionizing how we diagnose, manage, and prevent heart disease.The utilization of machine learning algorithms in CAHD prediction can have significant clinical implications. Early detection of individuals at high risk of developing CAHD allows for timely intervention and preventive measures. Personalized treatment strategies can be designed based on the predicted risk factors, optimizing patient care and improving health outcomes.

In the comparative study, KNN approach performed better performance and obtained anprecission score of 87.7%, recall rate of 82.1% and F1-score of 84.8% with CHAD dataset, compareed with other approaches. At the outset, machine learning algorithms offer a valuable tool for CAHD prediction in the early stage itself, i.e., it offers the medical practionarsto analyze complex patterns respect toCAHD feature ratings and identify high-risk individuals in the primitive stage, so healthcare professionals can assist to make informed valuable predictions and implementing targeted interventions in early state.

## REFERENCES

1. Noor Basha, Ashok Kumar P S, P Venkatesh, “ Early detection of heart syndrome using machine learning technique”, 2019, 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), PP 387-391

2. DAPS Prasadgouda B Patil, Dr. P Mallikarjun Shastry, MACHINE LEARNING BASED ALGORITHM FOR RISK PREDICTION OF CARDIO VASCULAR DISEASE (CVD), 2020, Journal of Critical Reviews 7 (9), pp 836-844

3. DPMMS Dr. Ashokumar P S, Prasadgouda B Patil, Heart Attack Detection Based On Mask Region Based Convolutional Neural Network Instance Segmentation and Hybrid Classification Using Machine Learning Techniques, 2021, Turkish Journal of Computer and Mathematics Education 12 (09), pp 2228– 2244

4. S Patil, PSA Kumar, P Patil, P Palagi, Data analysis & classification methodology for knowledge discovery in big data, 2017, International Conference on Inventive Systems and Control (ICISC)

5. Ashokumar P S, Prasadgouda B Patil , Dr. Vijay Bhaskar Reddy, “Prediction of Cardiovascular Diseases by Integrating Electrocardiogram (ECG) and Phonocardiogram (PCG) Multi-Modal Features using Hidden Semi Morkov Model”, 2022, International

Journal on Recent and Innovation Trends in Computing and Communication, Vol 10, Issue 10, PP 32-44

6. Noor Basha, K Manjunath, Mohan Kumar Naik, Ashok Kumar P S, P Venkatesh, Analysis and Forecast of Heart Syndrome by Intelligent Retrieval Approach, 2020, Intelligent Computing and Innovation on Data Science: Proceedings of ICTIDS

7. Ashok Kumar PS,Prasadgouda B Patil, Dr. P Mallikarjun Shastry, A Novel Approach for Prediction of Cardio Vascular Disease: An Improved Genetic Algorithm Approach Using Classifiers, 2020, International Journal of Advanced Science and Technology 29 (7), 4493 – 4504.

8. Ranjit Shrestha1 and Jyotir Moy Chatterjee, Heart Disease Prediction System Using Machine Learning, LBEF Research Journal of Science, Technology and Management, 2019, Vol 1, Issue 2,

9. M. Al Mehedi Hasan, J. Shin, U. Das, and A. Yakin Srizon, "Identifying Prognostic Features for Predicting Heart Failure by Using Machine Learning Algorithm," pp. 40–46, 2021, doi: 10.1145/3460238.3460245.

10. M. A. Khan et al., "Global Epidemiology of Ischemic Heart Disease: Results from the Global Burden of Disease Study," Cureus, vol. 12, no. 7, 2020, doi: 10.7759/cureus.9349.

11. Statista, "Small business statistics 'scary,'" Finweek, 2020. https://www.statista.com/statistics/880155/number-of-smes-in-unitedkingdom-uk/.

12. N. Kumar and D. Kumar, "Machine Learning based Heart Disease Diagnosis using Non-Invasive Methods: A Review," J. Phys. Conf. Ser., vol. 1950, no. 1, 2021, doi: 10.1088/1742-6596/1950/1/012081.

13. A. Haleem, M. Javaid, R. P. Singh, and R. Suman, "Applications of Artificial Intelligence (AI) for cardiology during COVID-19 pandemic," Sustain. Oper. Comput., vol. 2, no. February, pp. 71–78, 2021, doi: 10.1016/j.susoc.2021.04.003.

14. K. W. Johnson et al., "Artificial Intelligence in Cardiology," J. Am. Coll. Cardiol., vol. 71, no. 23, pp. 2668–2679, 2018, doi: 10.1016/j.jacc.2018.03.521.

15. A. Dudchenko, M. Ganzinger, and G. Kopanitsa, "Machine Learning Algorithms in Cardiology Domain: A Systematic Review," Open Bioinforma. J., vol. 13, no. 1, pp. 25–40, 2020, doi: 10.2174/1875036202013010025.