



Enhancing Myocardial Infarction Prediction Models via Hyperparameter Optimization

Tripti Thakur

Research Scholar, M.Tech.
Computer Science and Engineering
Takshshila Institute of Engineering and Technology
Jabalpur (M.P), India
Email: triptithakur@gmail.com

Swati Soni

Assistant Professor
Department of Computer Science and Engineering
Takshshila Institute of Engineering and Technology
Jabalpur (M.P), India
Email: swatisoni@takshshila.org

ABSTRACT

This study focuses on Enhancing Myocardial Infarction Prediction Models via Hyperparameter Optimization. Myocardial Infarction means heart attacks, is a critical health issue caused by the blockage of blood flow to the heart due to plaque buildup in the coronary arteries. Early detection and accurate prediction of heart disease are vital to prevent life-threatening situations. To achieve this, the study utilized a kaggle dataset comprising 303 rows and 14 columns, including features such as age, sex, cholesterol levels, and more. Five different machine learning models, namely Logistic Regression, Random Forest, K Neighbors Classifier, Decision Tree Classifier, and Support Vector Classifier, were employed to predict heart disease.

The models were initially evaluated using train-test split accuracy, with Logistic Regression achieving an accuracy of 91.80%, Random Forest at 88.52%, K Neighbors Classifier at 67.21%, Decision Tree Classifier at 72.13%, and Support Vector Classifier at 86.89%. Subsequently, to validate model performance, k-fold cross-validation ($k=10$) was applied, revealing mean accuracies of 83.83% for Logistic Regression, 83.46% for Random Forest, 62.39% for K Neighbors Classifier, 77.86% for Decision Tree

Classifier, and 83.83% for Support Vector Classifier.

Based on the cross-validation results, the best-performing models were selected and further fine-tuned using GridsearchCV and RandomsearchCV for hyperparameter optimization. The final results demonstrated improved performance:

- *Logistic Regression achieved an accuracy of 91.80% on the test data with an AUC-ROC score of 0.92.*
- *Random Forest exhibited exceptional accuracy, achieving 100% on the test data with an AUC-ROC score of 1.00.*
- *Support Vector Classifier achieved an accuracy of 93.44% on the test data with an AUC-ROC score of 0.93.*

This study highlights the importance of hyperparameter tuning in optimizing heart disease prediction models, with Random Forest and Support Vector Classifier showing promising results for accurate heart disease prediction.

Keywords:— Myocardial Infarction, ECG, Machine Learning, k-fold Cross Validation, Hyperparameter Tuning, AUC-ROC Score.

I. INTRODUCTION

Myocardial Infarction

When there is a significant reduction or obstruction in the blood supply to the heart, Myocardial

Infarction also known as heart attack happens. The accumulation of fat, cholesterol, and other materials in the heart's (coronary) arteries is typically the cause of the obstruction. The deposits that are high in fat and cholesterol are referred to as plaques. Atherosclerosis is the term for the plaque accumulation process. A plaque may occasionally burst and create a clot that stops blood flow. A portion of the heart muscle may get harmed or destroyed by inadequate blood flow.

An artery that supplies the heart with blood and oxygen becomes clogged, which leads to a heart attack. Over time, deposits of fat and cholesterol accumulate in the heart's arteries to create plaques. A blood clot may occur if a plaque breaks. The clot may clog arteries, leading to heart problems.

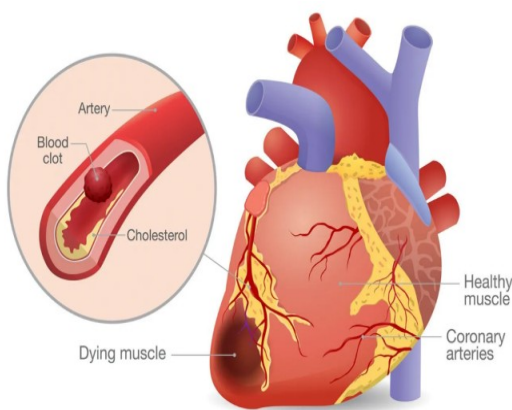


Figure 1. Blood Clot in Coronary Artery

Myocardial Infarction Symptoms: Symptoms of a heart attack vary. Some people have mild symptoms. Others have severe symptoms. Some people have no symptoms. Common heart attack symptoms include:

Chest pain that may feel like pressure, tightness, pain, squeezing or aching

Pain or discomfort that spreads to the shoulder, arm, back, neck, jaw, teeth or sometimes the upper belly

- Cold sweat
- Fatigue
- Heartburn or indigestion
- Lightheadedness or sudden dizziness
- Nausea
- Shortness of breath

Women may have atypical symptoms such as brief or sharp pain felt in the neck, arm or back. Sometimes, the first symptom sign of a heart attack is sudden cardiac arrest.

Some heart attacks strike suddenly. But many people have warning signs and symptoms hours, days or weeks in advance. Chest pain or pressure (angina) that keeps happening and doesn't go away with rest may be an early warning sign. Angina is caused by a temporary decrease in blood flow to the heart.

Range of Chest Pains: Angina: Angina is a type of chest pain or discomfort that is frequently brought on by a restricted or blocked coronary artery, which reduces blood flow to the heart muscle. Angina is described as a pressing or squeezing sensation in the chest. It happens when the heart isn't receiving enough blood. In addition, a person's shoulder, back, neck, arms, and mouth may hurt.

Typical Angina: Usually brought on by physical activity or mental stress, typical angina is a particular kind of chest pain that is eased by rest or nitroglycerin. It is characterized by a tight, squeezing sensation in the chest.

Atypical Angina: Pain or discomfort in the chest that does not follow the traditional

pattern of typical angina is referred to as atypical angina. It could have diverse characteristics and not always react to nitroglycerin or be induced by physical activity.

Non-Anginal Pain: Pain or discomfort in the chest that is unrelated to coronary artery disease is referred to as non-anginal pain. It can be caused by a number of things, including respiratory disorders, gastrointestinal disorders, anxiety, and musculoskeletal problems.

Asymptomatic: This refers to the absence of any discernible symptoms. It suggests that there is no discomfort or pain in the chest when it comes to chest pain.

Electrocardiogram (ECG): The ECG captures the electrical signal generated by the heart in order to detect various cardiac disorders. The electrical impulses that cause the heart to beat are recorded using electrodes applied to the chest. The signals appear on a connected printer or computer monitor as waves.

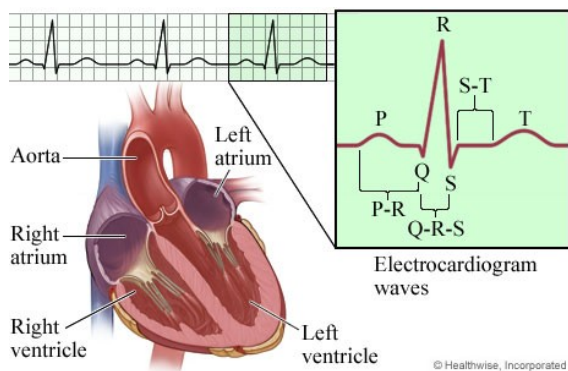


Figure 2. Representation of P-Wave QRS-Complex and T-Wave

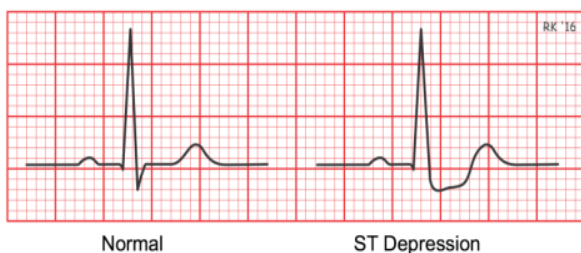


Figure 3. ECG ST Depression

II. MACHINE LEARNING

Within the field of artificial intelligence (AI), machine learning gives computers the ability to learn and make predictions or judgments based on information without the need for explicit programming. Although it includes a variety of methods, the three main kinds are as follows:

Supervised Learning: Using labeled datasets, algorithms are trained with input data coupled with matching target outputs. To enable the model to produce precise predictions for fresh, unknown data, the goal is to learn a mapping from inputs to outputs. Unsupervised Learning: The aim of unsupervised learning is to find structures, correlations, or patterns in unlabeled datasets. Typical assignments include decreasing the dimensionality of the data or grouping together comparable data points.

Reinforcement Learning: The goal of reinforcement learning is to teach algorithms to make choices in a way that maximizes cumulative rewards over an extended period of time. Through interaction with the environment and feedback in the form of incentives or punishments for its actions, the model gains knowledge.

By enabling computers to improve over time through experience, machine learning adds to artificial intelligence. Within machine learning, there are two primary categories of tasks:

Classification: As a type of supervised learning, classification entails classifying input data points into distinct groups or categories. The model is appropriate for jobs where the output variable is categorical since it learns to map input data to specified output classes. Regression: Predicting continuous numerical values from input data is the goal of regression, another kind of supervised learning. Regression tasks are

beneficial when the output variable is continuous because the model learns to create a link between the input features and the output variable.

In conclusion, regression is used to predict continuous numerical values, whereas classification is utilized to deal with categorical output factors. These two core categories of machine learning tasks are essential to many different kinds of applications in many different fields.

Classification is the proper method to handle this issue given the nature of the AHD Column in Dataset “heart.csv”. Machine Learning Classification Algorithms used Logistic Regression, Random Forest, K Neighbors, Decision Tree, and Support Vector Classifier

k-fold Cross Validation: Splitting tests into trains has drawbacks. First off, because the selection of data points for the training and testing sets is arbitrary, it may lead to a large degree of variability in performance evaluation. Because of this randomness, a model’s performance might range dramatically between train-test splits, making it difficult to make reliable conclusions about how well a model generalises to new data. Furthermore, data waste may result from train-test splitting, particularly in cases where the dataset is limited. There is less data available for training since some of it is set aside for testing. Models with fewer training data may be less resilient or more prone to over fitting, which is particularly troublesome when working with small amounts of data. We use k-Fold Cross-Validation to solve these problems. With this method, the data is divided into k folds, or subsets, and each fold is used iteratively as training and testing set. An estimate of a model’s capacity for generalisation is produced that is more reliable and representative by averaging the performance over these

rounds. Additionally, cross-validation guarantees effective data use by giving every data points a chance to be included in both the training and testing sets. This is especially important when there is a lack of data. Cross-validation is a solid method for evaluating and enhancing machine learning models since it helps with model selection and hyperparameter tweaking and allows for trustworthy comparisons across various models or parameter settings.

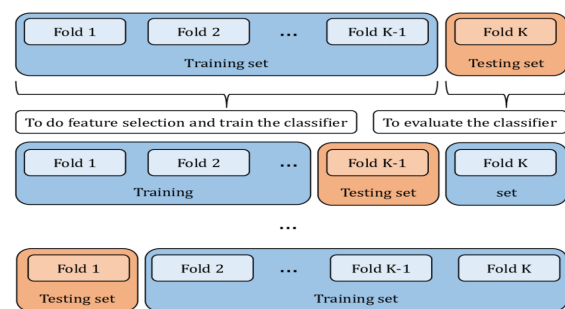


Figure 4. k-fold Cross Validation

Hyperparameter Tuning: Hyperparameter tuning, also known as hyperparameter optimization, is the process of systematically searching for the best combination of hyperparameters for a machine learning model to achieve optimal performance. Hyperparameters are external configuration settings that govern how a machine learning algorithm learns and generalizes from the training data. Unlike model parameters, which are learned from the data during training (e.g., weights in a neural network), hyperparameters are set by the data scientist or machine learning practitioner before training begins.

The objective of hyperparameter tuning is to find the hyperparameter values that result in the best model performance, such as maximizing predictive accuracy, minimizing error, or achieving a specific performance metric. This process typically involves trying different values or ranges for hyperparameters, training the model with each set of hyperparameters, and

evaluating the model’s performance on a validation dataset. Techniques for hyperparameter tuning include grid search, random search, Bayesian optimization, and more, depending on the complexity of the problem and available computational resources. The ultimate goal of hyperparameter tuning is to fine-tune the model, prevent overfitting, and improve its generalization ability to make accurate predictions on unseen data. GridsearchCV and RandomizedSearchCV are two popular techniques for hyperparameter tuning in machine learning. They both aim to find the best combination of hyperparameters for a model, but they use different strategies for exploring the hyperparameter space. GridSearchCV(Grid Search Cross-Validation):

GridsearchCV is an exhaustive search technique that evaluates the model’s performance for all possible combinations of hyperparameters within a predefined range or set. It creates a grid of hyperparameter values and trains and validates the model for each combination. This method is systematic and ensures that you test every specified combination of hyperparameters. However, it can be computationally expensive, especially when dealing with a large number of hyperparameters or a wide range of values.

RandomSearchCV (Randomized Search Cross -Validation): RandomizedSearchCV, on the other hand, takes a more randomized approach. Instead of evaluating all possible combinations, it randomly samples a specified number of hyperparameter

Data element	Description	Type	Range	Remarks
Age	-	Num ^a	29-77	Average is 54.37
Sex	-	Bi ^b	0: Female 1: Male	32% Female 68% Male
Cp	Chest pain level	Nom ^c	0/1/2/3 0: Asymptotic 2: non-anginal pain 3: Typical angina	Majority have 0 pain
Trestbps	Rest blood pressure	Num	94-200	Average is 131.6
Chol	Cholesterol level	Num	126-564	Average is 246.3
Fbs	Fasting blood sugar level	Bi	0: Level below 120 1: Level above 120	-
Restecg	Resting electrocardiographic results	Nom	0/1/2 0: Showing probable or definite left ventricular hypertrophy. 2: Abnormal	-
Thalach	Maximum heart rate achieved	Num	71-202	-
Exang	Exercise induced angina	Bi	0: None 1: Produced	-
Oldpeak	ST depression induced by exercise relative to rest	Num	0-6.2	Right skewed data, majority of population is between 0 and 0.5
Slope	The slope of the peak exercise ST segment	Nom	0: Unslowing 1: Flat 2: Down-sloping	-
Ca	Number of major vessels	Nom	0/1/2/3/4	-
Thal	Defect type	Nom	1: Fixed defect 2: Normal 3: Reversible defect	There is one outlier of category 0
Target	Diagnosis of heart disease	Bi	0: No disease 1: Disease	-

Figure 5. Heart Disease Dataset Description

configurations from the hyperparameter space. This randomness makes it more computationally efficient, especially when the hyperparameter space is extensive. RandomizedSearchCV can quickly explore a wide range of hyperparameters, potentially finding good configurations faster than GridsearchCV. However, there's a chance it may miss the optimal combination due to its randomized nature.

III. LITERATURE REVIEW

This work dataset was collected from Kaggle. The dataset contains a total of **303 instances with 13 attributes**.

Data Preprocessing: Data preparation is crucial because the caliber of the data used to construct a machine learning model heavily influences its performance. Data preprocessing comprises converting, resampling, and feature selection in addition to cleaning the data by eliminating outliers and corrupted or missing data points[1].

Attributes	Outlier values
Age	None
Chol	417, 564, 394, 407, 409
Trestbps	172, 178, 180, 180, 200, 174, 192, 178, 180
Thalach	71
Oldpeak	4.2, 6.2, 5.6, 4.2, 4.4

Figure 6. List of Outliers

Only the extreme outliers were eliminated because the mild outliers help determine the final diagnosis. (1) & (2) were used to identify the extreme outliers. In these formulas, Q1, Q3 represent the lower and upper quartiles, respectively, while IQR stands for interquartile range, which measures the dispersion of the data.

$$(75\% \times Q3) + 3 \times IQR \dots\dots\dots(1)$$

$$(25\% \times Q1) - 3 \times IQR \dots\dots\dots(2)$$

Because the mild outliers contribute to the final diagnosis, only the extreme outliers

removed were the data points that exceeded the first expression. In a similar manner, we eliminated the data points that are smaller than the second expression. Consequently, two of the 303 cases were eliminated. "IQR" is an acronym meaning "Interquartile Range." A statistical tool for evaluating the distribution or dispersion of data points within a dataset is the interquartile range. The difference between the first quartile (Q1) and the third quartile (Q3) is used to compute it.

Here's the formula for calculating the interquartile range (IQR):

$$IQR = Q3 - Q1$$

Where:

Q1 is the first quartile, representing the 25th percentile of the data.

Q3 is the third quartile, representing the 75th percentile of the data.

When attempting to comprehend the variability in the middle 50% of the data, the IQR is a helpful metric. It is frequently used to find data points that considerably deviate from this middle range, which are regarded as possible outliers, in conjunction with box plots and outlier identification techniques.

As mentioned in the previous comment, the IQR is utilized in the context of the information supplied to establish thresholds for finding severe outliers in particular dataset properties [1].

The work focused on predicting heart disease in patients using machine learning algorithms, specifically Support Vector Machine (SVM) and K-Nearest Neighbor (KNN).

The primary goal of this research is to predict which patients are more likely to suffer from heart disease based on various

medical features, such as chest pain, extracted from their medical records. This prediction is essential for early intervention and treatment. The dataset used for this research is obtained from Kaggle.

Since the dataset contains no missing values, outliers, or categorical data, the only preprocessing step performed is feature selection. Machine Learning Algorithms used, Support Vector Machine (SVM) and

K-Nearest Neighbor (KNN) Libraries Used: Several Python libraries are utilized, including the Math Library, NumPy, Pandas, Plotly, and Matplotlib, for various purposes such as data manipulation, visualization, and calculations. Ec: This represents the number of samples that were expected to be classified as positive (in the context of predicting heart disease). Enc: This represents the number of samples that were expected to be classified as negative (not having heart disease) [2].

Total Samples =	Output		
103	Expected output	Ec	Enc
c=55 positive		41	14
nc = 48 negative		3	45

Figure 7. Result & Analysis of SVM

Total Samples =	Output		
257	Expected output	Ec	Enc
c=123 positive		101	22
nc = 134 negative		13	121

Figure 8. Result & Analysis of KNN

Algorithm	Accuracy %	Miss rate %	Precision %	Recall %	F1 Score %
KNN	86	14	84	90	87
SVM	83	16	76	94	84

Figure 9. Result & Analysis of KNN, SVM

The principal objective of the work is to utilize machine learning, more especially the Random Forest method, to precisely forecast the probability of heart disease, an important issue in contemporary medicine. Collection: Age, the type of chest pain (Cp), resting blood pressure (Trestbps), cholesterol levels (Chol), fasting blood sugar (Fbs), maximum heart rate achieved (Thalach), exercise-induced angina (Exang), ST depression caused by exercise relative to rest (Old Peak), and the type of Thalassemia (Thal) are among the patient attributes included in the dataset. Additionally, a “Target” column is included, which indicates if heart disease is present (1) or absent (0). Algorithm: The Random Forest machine learning algorithm was selected because of its reputation for effectively handling both regression and classification tasks. In order to improve prediction accuracy, it works by building an ensemble of decision trees and combining their results. Instruction and Assessment Dividing: The dataset is split into two subsets in order to assess the model’s performance: a training set, which normally makes up 70% of the data, and a testing set, which makes up the remaining 30%. The testing set is used to evaluate the model’s accuracy and generalization to new, unobserved data, whereas the training set is used to educate the model to generate predictions.

Accuracy: The emphasis throughout the work is on achieving high performance and accuracy rates. The accuracy of the prediction model depends on various factors, including the number of decision trees in the Random Forest. By combining multiple classifiers, the algorithm strives to increase accuracy while avoiding overfitting [3].

Disease prediction systems are highlighted as valuable tools for avoiding human errors in disease diagnosis and assisting in early

disease prevention. The research focuses on developing intelligent heart disease prediction systems. The dataset employed in this study is the renowned Cleveland heart disease dataset, sourced from the **UCI machine learning repository**. Comprising a total of **303 heart disease data records**, this dataset offers a comprehensive set of **76 attributes**, each providing unique insights into various patient health-related factors. However, to focus specifically on heart disease prediction, the research selects a subset of **14 attributes deemed most relevant for the task**. These attributes encompass a diverse range of information, including patient age, gender, chest pain type, blood pressure, cholesterol levels, and more. The dataset contains a mix of data types, including integers, real numbers, and discrete values, reflecting the varied nature of health-related information. Crucially, the “num” attribute within the dataset serves as the output label, with “num=0” indicating normal health and “num=1” signifying the presence of heart disease.

Platforms, including Weka, Rapid Miner, Mahout, and MATLAB [4].

Results – 80.4% of decision trees, 85.68% are Neural Networks, 86.12% using Naïve Bayes

Genetic Algorithm Combined with Hybrid Neural Network:

Accuracy of training dataset: 96.2%

92% of the test dataset is accurate.

Accuracy of validation set: 89%

Sturdy Intelligent Heart Disease Prediction System (RIHDPS): 91.26% of predictions are accurate.

89.32% is the Naïve Bayes accuracy value.

Accuracy percentage for logistic regression: 84.47%

Accuracy value for neural networks: 90.2%

Artificial Neural Network (ANN) and K-Means Clustering Hybrid Model: 93.5% [4].

Study focused on the diagnosis and prediction of heart-related diseases using machine learning algorithms. The introduction highlights the significance of the heart in living organisms and the importance of accurate diagnosis and prediction of heart-related diseases due to the potential life-threatening consequences. The main objective of the work is to develop a prediction system for heart diseases using machine learning algorithms to raise awareness about these diseases.

Data Collection: The study uses a dataset from the UCI repository, which is widely recognized and verified.

Attribute Selection: Various attributes from the dataset are chosen for the prediction system, such as age, gender, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, etc.

Data Preprocessing: It is mentioned as a crucial step to achieve accurate results from machine learning algorithms. It includes handling null values and converting categorical variables into numerical form using dummy encoding (0 and 1).

Machine Learning Algorithms [5]: The paper evaluates the performance of several machine learning algorithms for predicting heart disease, Support Vector Machine (SVM), Decision Tree, Linear Regression, K-nearest Neighbor (KNN).

The paper presents the accuracy of each machine learning algorithm in predicting heart disease based on the dataset. The

accuracy percentages are provided for each algorithm:

- Support Vector Machine: 83%
- Decision Tree: 79%
- Linear Regression: 78%
- K-nearest Neighbor: 87%

IV. METHODOLOGY

This research aims to contribute to the development of accurate and accessible heart disease prediction systems, ultimately improving healthcare outcomes and patient awareness. Accurate prediction of myocardial infarction is a critical healthcare objective, and traditional methods like train-test splitting have shown limitations in delivering precise results. This research focuses on elevating heart disease prediction models to new heights by implementing advanced model validation techniques and hyperparameter optimization. The primary aim is to enhance prediction accuracy and reliability for effective early intervention and treatment.

Cross Validation: Apply k-fold cross-validation on the (x, y) data to evaluate model performance across diverse validation sets. **Hyperparameter Tuning:** Identify the best-performing models based on cross-validation accuracy. Fine-tune selected models through GridsearchCV and RandomsearchCV to optimize hyperparameters.

Data Source: Dataset - <https://www.kaggle.com/datasets/zhaoyingzhu/heartcsv>,

File Size : 19KB, Dataset Size: 303 rows × 14 columns

As All Values of AHD Column are either 0 or 1 Thus, it is a Classification Problem.

Proposed-Framework:

- IDE - Google Collaborator
- Python–Python 3
- Applied Binary Classification Problem on Data

Importing Libraries

Drive Mount & Data Collection: Reading CSV file (heart.csv) using pandas

Data Preprocessing Starts: Printing dataframe df / data Exploration to obtain information about a DataFrame

Check for Null Values: Checking for missing values in the dataframe using `df.isnull().sum()`.

Fill Null Values with Mean

Data Preprocessing (Continued):

- Separating the Dataset Features into features (x) and target (y).
- get the Unique Values of Columns
- Plot Pie Chart for target (y)
- Standardization is used to bring features onto a common scale

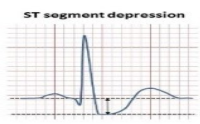
Further splitting the data into training and testing sets using `train_test_split` with 80% training and 20% testing data.

Model Creation:

Classification models are created on Training – Testing Data and evaluated using Confusion Matrix. The models are:

- Logistic Regression Model
- Random ForestClassification Model
- K Neighbors Classifier Model
- Decision Tree Classifier Model
- Support Vector Classifier Model

Table 1. Heart Disease Prediction Dataset

Sr. No.	Feature	Description	Detailed Description
1	Age	The person's age in years	Type: Numerical Range: 29-77
2	sex	The person's sex (1 = male, 0 = female)	Type: Binary 68% Mal, 32% Female
3	cp:	chest pain type	— Value 0: asymptomatic — Value 1: atypical angina — Value 2: non-anginal pain — Value 3: typical angina
4	restbps:	The person's resting blood pressure (mm Hg on admission to the hospital) Dataset Value:94-200	Blood pressure is measured in millimetres of mercury (mmHg) and has 2 types: Systolic pressure – the pressure when your heart pushes blood out. Diastolic pressure – the pressure when your heart rests between beats. Normal Range: 90-120
5	chol:	The person's cholesterol measurement in mg/dl Dataset Value: 126- 564	Cholesterol is usually measured in milligrams (mg) of cholesterol per deciliter (dL) of blood. Healthy Level Cholesterol is 125 to 200mg/dL
6	fbs:	The person's fasting blood sugar Dataset Value: (> 120 mg/dl, 1 = true; 0 = false)	The expected values for normal fasting blood glucose concentration are between 70 mg/dL and 100 mg/dL.
7	restecg:	resting electrocardiographic results Values: 0,1,2	Value 0: showing probable or definite left ventricular hypertrophy by Estes' criteria Value 1:Normal Value 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
8	thalach:	The person's maximum heart rate achieved. Values: 71 - 202	Normal: 60 to 100Per Minute
9	exang:	Exercise induced angina (1 = yes; 0 = no)	Chest Pain after Exercise
10	oldpeak:	ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot.) 1 Block Size – 1mm* 1mm Block Inside Block – 0.2 mm	
11	slope:	The slope of the peak exercise ST segment	0: downsloping; 1: flat; 2: upsloping
12	ca:	The number of Major Vessels (0–3) Blockage Severity 0/ 1/ 2/ 3	Coronary Arteries (Blood vessels -supply blood to the heart muscle)
13	thal:	A blood disorder called Thalassemia Value 0: NULL (dropped from the dataset previously) Value : 1 / 2/ 3	Thalassemia an inherited blood disorder that causes your body to have less hemoglobin than normal. When Hemoglobin unable to carry O2. Value 1: fixed defect (no blood flow in some part of the heart) Value 2: normal blood flow Value 3: reversible defect (a blood flow is observed but it is not normal)
14	AHD	Heart disease	1 = YES, 0= NO

V. RESULTS

Results of Train –Test Splitting, CV, Hyperparameter Optimization

Table 2. Proposed Work Results with various techniques

Classifier Models	Train Test Split	Cross Validation CV=10		GridsearchCV /RandomsearchCV
Logistic Regression	Accuracy - 0.9180327868852459	[0.90322581 0.80645161 0.80645161 0.96666667 0.83333333 0.7 0.86666667 0.9 0.73333333 0.86666667]	mean_accuracy_lr 83.828	Best Hyperparameters from Grid Search: {'C': 0.1, 'penalty': 'l2'} Best Accuracy: 0.8449462365591398 Best Hyperparameters from Random Search: {'penalty': 'l2', 'C': 0.1} Best Accuracy: 0.8449462365591398
Random Forest	Accuracy – 0.8852459016393442	[0.90322581 0.83870968 0.87096774 0.93333333 0.9 0.8 0.7 0.83333333 0.73333333 0.83333333]	mean_accuracy_RF 83.4624	Best Hyperparameters from Grid Search: {'max_depth': 10, 'n_estimators': 200} Best Accuracy: 0.8281720430107526 Best Hyperparameters from Random Search: {'n_estimators': 200, 'max_depth': 10} Best Accuracy: 0.8281720430107526
KNeighbour	Accuracy - 0.6721311475409836	[0.70967742 0.64516129 0.48387097 0.66666667 0.6 0.5 0.66666667 0.7 0.53333333 0.73333333]	mean_accuracy_KNC 62.3871	
Decision Tree	Accuracy - 0.819672131147541	[0.77419355 0.87096774 0.77419355 0.86666667 0.76666667 0.76666667 0.76666667 0.76666667 0.7 0.73333333]	mean_accuracy_DT 77.8495	
Support Vector	Accuracy - 0.8688524590163934	cv_score_svc [0.87096774 0.80645161 0.83870968 0.96666667 0.83333333 0.7 0.86666667 0.9 0.70.9]	mean_accuracy_svc 83.828	Best Hyperparameters from Grid Search: {'C': 0.1, 'kernel': 'linear'} Best Accuracy: 0.8447311827956989 Best Hyperparameters from Random Search: {'kernel': 'linear', 'C': 0.1} Best Accuracy: 0.8447311827956989

Table 3. Final Model Results After Applying Best Parameters

Sr.	Machine Learning Model	Trained & Tested On	Confusion Matrix	Accuracy on Test Data	AUC-ROC Score
1.	Logistic Regression	fit(x,y), Predict(x_test)	[[30 3] [2 26]]	0.9180327868852459	0.92
2.	Random Forest	fit(x,y), Predict(x_test)	[[33 0] [0 28]]	1.0	1.00
3.	Support Vector Classifier	fit(x,y), Predict(x_test)	[[31 2] [2 26]]	0.9344262295081968	0.93

Final Model Creation with Best Parameters:

- Logistic Regression Model-Accuracy on Test Data: 0.9180327868852459

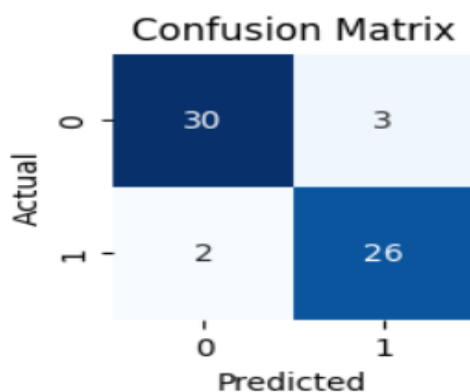


Figure 10. Logistic Regression Model with Best Parameters

- Random Forest Classification Model-Accuracy on Test Data: 1.0

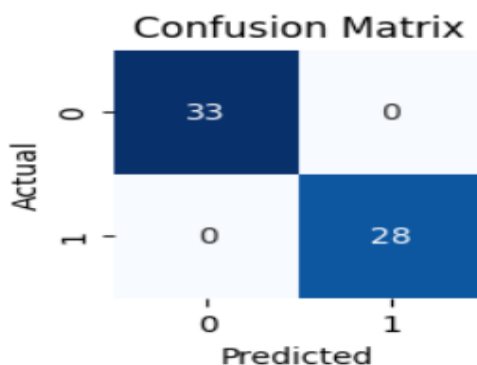


Figure 11. Random Forest Classification Model with Best Parameters

- Support Vector Classifier Model-Accuracy on Test Data 0.9344262295081968

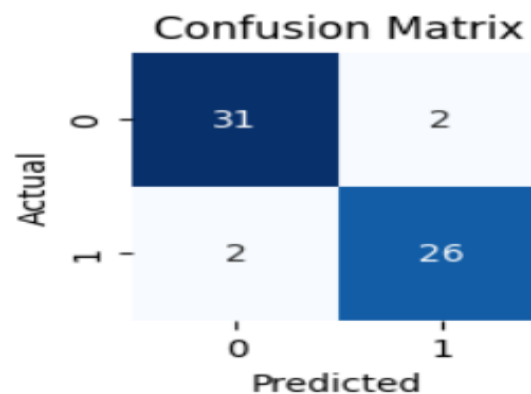


Figure 12. Support Vector Classifier Model with Best Parameters

VI. CONCLUSIONS AND FUTURE WORK

Conclusions

In this study, we conducted a thorough performance evaluation of various machine learning algorithms for the prediction of heart disease. Our methodical technique comprised of train-test splitting, k-fold cross-validation (CV=10), and GridsearchCV and RandomsearchCV for hyperparameter tweaking. The outcomes offered insightful information on how well these methods work to maximize model performance.

Train-Test Splitting: We divided the dataset into an 80-20 train-test split, in which 20% was set aside for testing the

model's prediction power and the remaining 80% was utilized for training the models. This method made sure that the models were evaluated on untested data, which aided in determining how well they generalized.

Cross-Validation: The dataset was split into ten subsets using k-fold cross-validation with k=10. The models were rotated through each subset as a validation set and then trained and evaluated ten times, using the remaining nine subsets for training. An assessment of the model's prediction accuracy that was more accurate was given by this robust cross-validation method. Each model's stated mean accuracy numbers provide a thorough evaluation of how well it performed across several validation sets. **Hyperparameter Tuning:** The model's hyperparameters were adjusted using GridsearchCV and RandomsearchCV. Through a thorough search, the optimal hyperparameters for each algorithm were found, maximizing performance. For every model, the optimal accuracy scores obtained through hyperparameter adjustment were noted.

Based on the maximum accuracy following hyperparameter adjustment, the Random Forest model was the best. The Random Forest model obtained 100% accuracy on the test dataset after using the optimal hyperparameters {'max_depth': 10, 'n_estimators': 200}. This shows that, after hyperparameter adjustment, the Random Forest model outperformed all the other models you assessed in your research, achieving flawless accuracy on the test data.

Future Work

- **Ensemble Learning:** To integrate the advantages of several models, investigate ensemble learning strategies like bagging and boosting. There may be more room for

improvement in prediction accuracy with ensemble techniques like Gradient Boosting and AdaBoost.

- **Feature Engineering:** Look into more sophisticated feature engineering methods to lower dimensionality or provide new, pertinent features. To improve model performance, take into account feature selection methods and domain-specific feature engineering.
- **Data Augmentation:** To enhance model generalization, collect a bigger and more varied dataset. The models may function even better and be able to handle a wider range of patient profiles if the dataset is expanded.
- **External Validation:** To make sure the models generalize far beyond the original dataset, validate the models using independent, external datasets from various healthcare facilities or geographical areas.
- **Real-time Monitoring:** Create a system for monitoring data in real-time so that models are updated with fresh information on a regular basis. This will enable the models to adjust to the changing patient demographics and changing medical procedures.
- **Patient Education:** Using the model projections, create educational resources or applications to educate patients about their heart disease risk factors. Promote proactive lifestyle modifications and healthcare practices.
- **Other Algorithms:** Investigate other deep learning architectures or machine learning algorithms to determine whether they can outperform the existing models. Try out more sophisticated methods like gradient-boosted trees or neural networks.

REFERENCES:

- [1] Chaimaa Boukhatem, Heba Yahia Youssef, Ali Bou Nassif, "Heart Disease Prediction Using Machine Learning", 2022 Advances in Science and Engineering Technology International Conferences (ASET) | 978-1-6654-1801-0/22/\$31.00 ©2022 IEEE | DOI: 10.1109/ASET53988.2022.9734880.
- [2] Taher M. Ghazal, Amer Ibrahim, Ali Sheraz Akram, Zahid Hussain Qaisar, Sundus Munir, Shanza Islam, "Heart Disease Prediction Using Machine Learning", 2023 International Conference on Business Analytics for Technology and Security (ICBATS) | 979-8-3503-3564-4/23/\$31.00 ©2023 IEEE | DOI: 10.1109/ICBATS57792.2023.10111368.
- [3] M.Snehith Raja, M.Anurag, Ch.Prachetan Reddy, Nageswara Rao Sirisala, "Machine Learning Based Heart Disease Prediction System", 2021 International Conference on Computer Communication and Informatics (ICCCI) | 978-1-7281-5875-4/21/\$31.00 ©2021 IEEE | DOI: 10.1109/ICCCI50826.2021.9402653.
- [4] Dr. Lakshmi Prasad Koyi, Tejaswi.Borra, Dr. G. Lakshmi Vara Prasad, "021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/\$31.00 ©2021 IEEE | DOI: 10.1109/ICAIS50930.2021.9395785".
- [5] Archana Singh, Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms", 2020 International Conference on Electrical and Electronics Engineering (ICEE-2020), 978-1-7281-5846-4/20/\$31.00 ©2020 IEE.
- [6] Vijeta Sharma, Shrinkhala Yadav, Manjari Gupta, "Heart Disease Prediction using Machine Learning Techniques", 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) | 978-1-7281-8337-4/20/\$31.00 ©2020 IEEE | DOI: 10.1109/ICACCCN51052.2020.9362842.
- [7] P. Ramprakash, R. Sarumathi, R. Mowriya, S. Nithyavishnupriya, "Heart Disease Prediction Using Deep Neural Network", Proceedings of the Fifth International Conference on Inventive Computation Technologies (ICICT-2020), IEEE Xplore Part Number:CFP20F70-ART; ISBN:978-1-7281-4685-0.
- [8] Rahul Katarya.Polipireddy Srinivas, "Predicting Heart Disease at Early Stages using Machine Learning: A Survey", Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020), IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281-4108-4.
- [9] Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Prof. Kailas Devadkar (PhD), "Prediction of Heart Disease Using Machine Learning", Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018), IEEE Conference Record # 42487; IEEE Xplore ISBN:978-1-5386-0965-1.
- [10] Prof. Kalpesh Joshi, Shubham Patil, Sagar Patil, Sahil A. Patil, Sahil S. Patil, Shantanu Patil, Saniya Patil, "Heart Disease Prediction using Machine Learning", International

- Journal for Research in Applied Science & Engineering Technology (IJRASET), ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538, Volume 11 Issue VI Jun 2023- Available at www.ijraset.com.
- [11] Prof. Sachin Sambhaji Patil, Vaibhavi Dhumal, Srushti Gavale, Himanshu Kulkarni, Shreyash Wadmalwar, “Heart Disease Prediction using Machine Learning”, International Journal of Scientific Research in Science and Technology, Print ISSN: 2, 395-6011 | Online ISSN: 2395-602X (www.ijrst.com), doi : <https://doi.org/10.32628/IJSRST229676>.
- [12] Goud Harsha Vardhan, Nallamilli Sneha Sisir Reddy, Dr. K.M. Umamaheswari, “Heart disease prediction using machine learning”, International Journal of Health Sciences, 6(S2), 7804–7813, <https://doi.org/10.53730/ijhs.v6nS2.6955>.
- [13] Apurb Rajdhan, Milan Sai, Avi Agarwal, Dundigalla Ravi, Dr. Poonam Ghuli, “Heart Disease Prediction using Machine Learning”, International Journal of Engineering Research & Technology (IJERT). ISSN: 2278-0181, Vol. 9 Issue 04, April-2020.
- [14] Rishabh Magar, Rohan Memane, Suraj Raut, “Heart Disease Prediction Using Machine”, Journal of Emerging Technologies and Innovative Research (JETIR), © 2020 JETIR June 2020, Volume 7, Issue 6, www.jetir.org (ISSN-2349-5162).
- [15] Prof. M. M. Bhajibhakare, Naeem Shaikh, Dipesh Patil, “Heart Disease Prediction using Machine Learning”, International Journal for Research in Applied Science & Engineering Technology (IJRASET), ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177, Volume 7 Issue XII, Dec 2019- Available at www.ijraset.com.

* * * * *