# Heart Disease identification Using Ensemble Modals

Ali Gohar
*Artificial Intelligence*
*Ghulam Ishaq Khan Institute Of Science and Technology*
Topi, Pakistan
u2022081@giki.edu.pk

Maqsood Ahmed
*Artificial Intelligence*
*Ghulam Ishaq Khan Institute Of Science and Technology*
Topi, Pakistan
u2022656@giki.edu.pk

***Abstract*—Cardiovascular illness is one of the world's leading causes of death, and both prevention and treatment depend on early identification and precise diagnosis. In order to increase prediction accuracy and offer insights into important heart disease contributing factors, this effort builds machine learning models for evaluating people's heart disease status using the UCI Heart Disease dataset. Additionally, this process uses group learning techniques.**

***Index Terms*—Classification, Machine Learning, Ensemble Learning, Heart Disease, and UCI Dataset.**

## I. Introduction

A major cause of death and morbidity, cardiovascular diseases (CVDs) continue to be a major global health concern. The World Health Organization (WHO) reports that cardiovascular diseases (CVDs) are the world's leading cause of death, accounting for 32% of all fatalities and 17.9 million deaths annually [1]. These problems include coronary artery disease, heart failure, arrhythmias, and other heart and blood vessel disorders. Ageing populations, genetic predispositions, and bad lifestyle choices are some of the reasons contributing to the rising prevalence of CVDs.

Reducing the impact of heart issues requires early detection and prompt action. Research indicates that prompt diagnosis enhances patient outcomes by facilitating more targeted treatment and preventative actions [2]. However, because traditional diagnostic approaches depend on invasive procedures or clinical skill, their effectiveness and accessibility may be limited. To improve the precision and effectiveness of heart disease detection, innovative, data-driven methods are therefore becoming more and more necessary.

Machine learning (ML) has revolutionized the medical field because it can evaluate vast amounts of complex data and spot patterns that are hard for humans to see. By integrating the predictive power of several models to increase accuracy and robustness, ensemble machine learning approaches in particular have demonstrated exceptional effectiveness in a variety of applications [3]. Because of their exceptional performance in classification problems, ensemble approaches like as Random Forest, gradient enhancement machines, and XGBoost are perfect for medical diagnostics [4].

This study looks into the application of ensemble learning techniques for the classification of heart sickness using the UCI Heart sickness dataset. This dataset, which contains a variety of patient data, including clinical, diagnostic, and demographic features, is widely known in the research community [5]. Using this information, the team aims to create incredibly accurate prediction models that provide insight into the primary causes of heart disease.

The following are this paper's main contributions:

1) The UCI Heart Disease dataset was analyzed to identify important characteristics that affect heart disease.
2) building ensemble machine learning models and evaluating how well they perform in comparison to traditional methods for classifying heart disease.
3) Clinical insights are obtained by an interpretability analysis that employs feature significance measures.
4) An examination of the real-world applications of machine learning in healthcare and how it might complement conventional diagnostic techniques.

## II. Dataset Description

The UCI Heart Disease dataset, which is commonly regarded as a benchmark in the machine learning area, is the source of the dataset used in this study. It consists of 1,025 training samples and 14 features, including both continuous and categorical variables. 'target' is a binary variable that indicates if a patient has heart disease (1) or not (0).

The following are the dataset's features:

- **age**: The age of person in years.
- **gender**: The gender of person (1 = male, 0 = female).
- **CP type)**: The types of chest pain in a person:
  - 0: Normal angina
  - 1: Abnormal angina
  - 2: Not anginal pain
  - 3: Asymptomatic
- **trestbps (resting BP)**:The resting blood pressure, expressed in millimeters of mercury (mm Hg), was obtained at the time of hospital admission.
- **chol (serum cholesterol)**: The cholesterol level in milligrams per deciliter.
- **fbs (fasting blood sugar)**: (if ¿120 mg/dL, 1; otherwise, 0) Fasting blood sugar.
- **restecg (resting electrocardiographic results)**:
  - 0: Normal
  - 1: Having ST-T wave abnormality
  - 2: Showing probable or definite left ventricular hypertrophy.

- highest heart rate attained while exercising is known as the ”thalach.”
- **exang (exercise-induced angina)**: Indicates whether angina was brought on by exercise (1 = yes, 0 = no).
- **oldpeak (ST depression)**: exercise-induced ST depression as compared to rest.
- **slope (slope of the peak exercise ST segment)**:
  - 0: Upsloping
  - 1: Flat
  - 2: Downsloping
- **ca (number of major vessels)**: Number of major vessels (0-3) colored by fluoroscopy.
- **thal (thalassemia)**:
  - 0: Normal
  - 1: Fixed defect
  - 2: Reversible defect
- **target**: The binary classification target (1 = heart disease present, 0 = no heart disease).

The dataset is appropriate for a range of machine learning approaches since it provides a well-balanced combination of continuous and categorical variables. There are enough observations in the training dataset to investigate trends and connections between the target variable and the characteristics. To maximize the performance of machine learning models, proper preprocessing is essential. This includes resolving missing values, encoding categorical variables, and normalizing data.

This URL will take you to the UCI heart disorders dataset. https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

## III. Data Preprocessing

To guarantee the dataset's quality and suitability for machine learning models, a number of preparation procedures were applied. To improve model performance, these procedures included cleaning, correcting missing values, renaming columns for clarity, encoding categorical variables, and discretizing continuous variables.

### A. *Renaming Categorical Variables*

The dataset was made easier to comprehend by giving categorical variables new, relevant labels.

- `Gender`: {0: *female*, 1: *male*}
- `CP_type`: {3: *asymptomatic*, 1: *atypical angina*, 2: *not anginal pain*, 0: *typical angina*}
- `fasting_blood_sugar`: {0: *less than 120mg/ml*, 1: *greater than 120mg/ml*}
- `rest_ecg`: {0: *normal*, 1: *ST-T wave abnormality*, 2: *left ventricular hypertrophy*}
- `EIA`: {0: *no*, 1: *yes*}
- `st_slope`: {0: *upsloping*, 1: *flat*, 2: *downsloping*}
- `thalassemia`: {1: *fixed defect*, 0: *normal*, 2: *reversible defect*}
- `condition (target)`: {0: *no disease*, 1: *has disease*}

### B. *Categorical and Continuous Variable Separation*

The dataset was separated into continuous and categorical variables to enable more efficient preprocessing:

- **Categorical Variables:** If a variable had less than or equal to ten different values, it was categorized as categorical.
- **Continuous Variables:** More than ten different values were considered continuous variables.

### C. *Handling Outliers*

Outliers were detected and removed using methods such as:

- **Isolation Forest:**An anomaly detection technique based on machine learning.
- **Elliptic Envelope:** A reliable technique to identify multivariate outliers.

The models' robustness and performance were enhanced by eliminating these outliers.

### D. *Discretization of Continuous Variables*

Continuous variables were discretized using the `KBinsDiscretizer`, transforming variables like `age`, `cholesterol`, and `st_depression` into binned categories to enhance interpretability and handle skewed distributions effectively.

### E. *Univariate Analysis*

**Categorical Data Observations:**

- The number of male observations was over twice that of female observations.
- The most common chest pain type was *asymptomatic*, accounting for nearly 50% of the data.
- 85% of patients did not have elevated fasting blood sugar levels. This can be shown in Figure 1

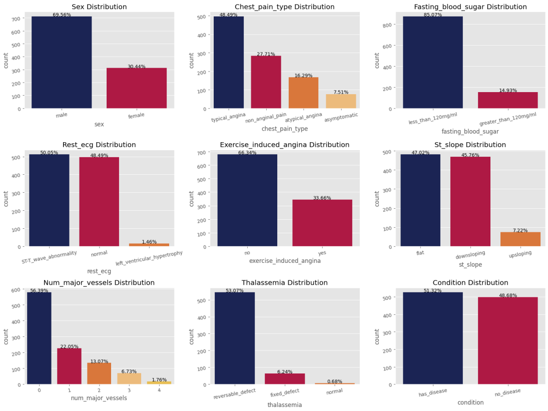

Fig. 1. Categorial Graphs.

**Continuous Data Observations:**

- MA Gaussian-like distribution with some skewness was observed for the majority of continuous variables.
- Outliers were found, particularly in the variables like `cholesterol`. This can be shown in Figure 2

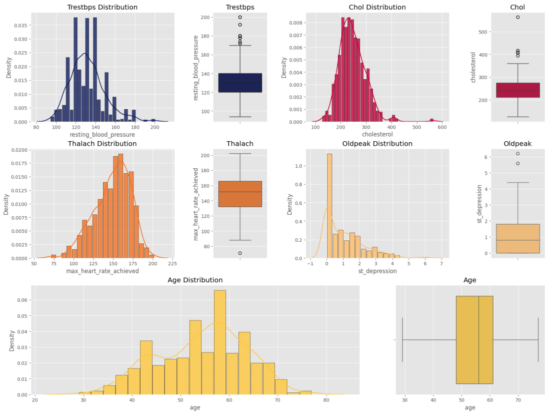

Fig. 2. Continuous Data Graph.

## IV. Bivariate Analysis

To investigate correlations between factors and the target variable, bivariate analysis was performed. (`condition`).

### A. Categorical Data vs Target

- When compared with women, men were far more probable to suffer from heart disease.
- Exercise-induced angina was an effective indicator of heart disease, with those who suffered it having a roughly threefold increased risk of developing the condition.
- Heart disease was more common among people with flat `st_slope` distributions. This can be shown in Figure 3

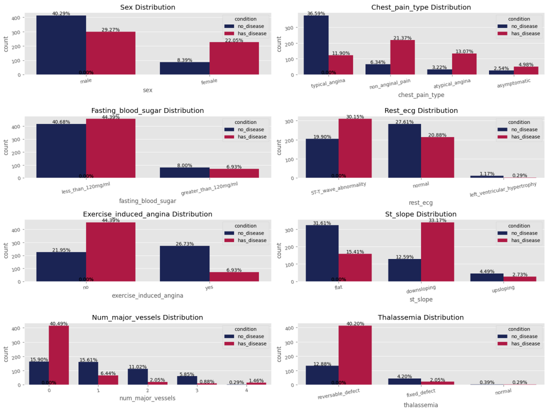

Fig. 3. Categorical Data vs Target.

### B. Continuous Data vs Target

- The risk of heart disease was marginally increased by increased resting blood pressure and cholesterol levels.
- The probability of heart disease increased with higher `st_depression` values.
- Heart disease was more common in older people. This can be shown in Figure 4

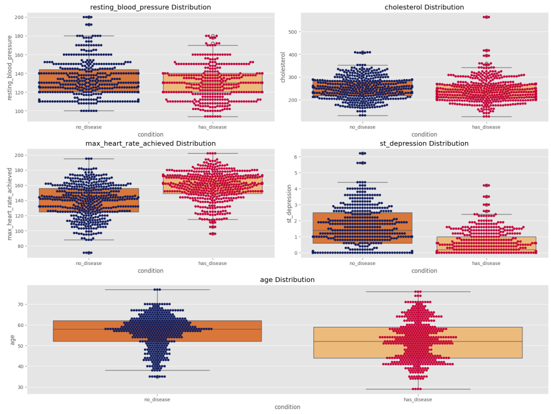

Fig. 4. Continuous Data vs Target.

## V. Correlation Analysis

To find linear relationships between variables, the Pearson correlation was used. The correlations were shown using a heatmap, which showed the following.

- Strong positive correlations between st$_{d}epressionandheartdisease.Weakornegligiblecorrelationsbetween$

## VI. Feature Importance

Models such as Random Forest and Gradient Boosting were used to determine the relevance of each feature. Among the most noteworthy characteristics were:

- `thalassemia`
- `st_depression (oldpeak)`
- `chest_pain_type (cp)`
- `exercise_induced_angina (exang)`
- `heart_rate_thalach`

To illustrate the influence of features on model predictions, feature significance visualizations were created. This can be shown in Figure 5

## VII. Modeling and Evaluation

### A. Models Implemented

The following classifiers were used to build predictive models:

- **Gradient Boosting Classifier:**
  The idea of boosting is expanded to work with any differentiable loss function via Gradient Tree Boosting, sometimes referred to as Gradient Boosted Decision Trees (GBDT). In a variety of domains, including environmental research and search engine ranking, this strong and adaptable machine learning technique performs exceptionally well in handling both regression and classification tasks.
- **Classifier for Random Forest:**
  The sklearn.ensemble module contains two randomized decision tree-based averaging algorithms: the RandomForest technique and the Extra-Trees approach. Both methods employ the perturb-and-combine method, which was developed specifically for trees. This suggests that a

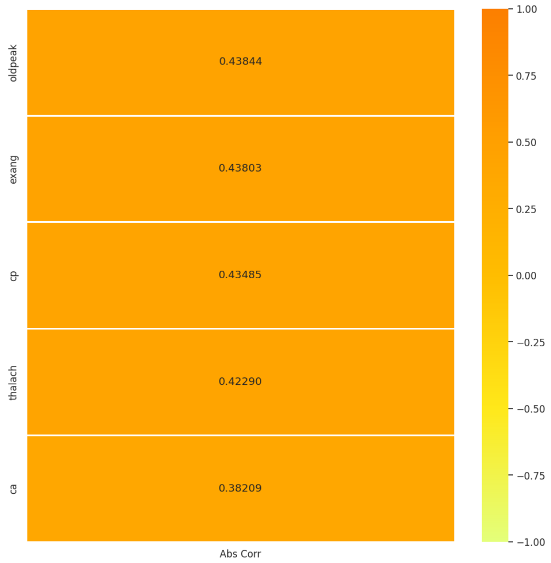

Fig. 5. Correlation Matrix.

diverse set of classifiers is produced when randomization is incorporated into the classifier design. The ensemble's prediction is based on the average forecast of each classifier.

- **AdaBoost Classifier**
  As a kind of meta-estimator, the AdaBoost classifier starts by using the original dataset to train a classifier. Then, using the same dataset and updated weights for the incorrectly categorized examples, it iteratively builds more classifiers. This modification guarantees that later classifiers concentrate more on the challenging cases.
- **SVM:**
  Support vector machines (SVMs) are a type of supervised learning methods used in outlier detection, regression, and classification. Some advantages of support vector machines include the following:
  effective in situations with several dimensions. It is still helpful when there are more dimensions than samples.
  Because it uses support vectors, a subset of training points in the decision function, it is also memory efficient. flexible: a range of Kernel functions can be used to produce the decision function. Custom kernels can be provided in addition to the regular kernels that are available.
- **MLP Classifier:**
  Stochastic gradient descent, also known as "Limited-memory Broyden-Fletcher-Goldfarb-Shanno" LBFGS, is used by the multi-layer Perceptron classifier to optimize the log-loss function.
- **KNN Classifier:**
  Neighbors-based classification is a kind of instance-based or non-generalizing learning that does not rely on creating a general internal model. Instead, it relies on storing the training examples. The approach determines the class of a new point by identifying its closest neighbors and voting by majority. The class with the highest representation among these neighbors is then given the query point.
- **DT Classifier:**
  A supervised learning method for classification and regression problems is a decision tree (DT). By creating a model using a set of straightforward if-then-else rules that are drawn from the data, it functions as a non-parametric approach. The main choices required to forecast the target variable are intended to be captured by these rules. A decision tree, for instance, can learn patterns from the data iteratively and approximate a sine curve. The model can match the data better with deeper trees since they produce more complex decision rules.
- **Gaussian Naive Bayes:**
  The Gaussian Naive Bayes classification algorithm is carried out by GaussianNB. It is assumed that the attributes have a Gaussian likelihood.

### *B. Hyperparameter Tuning*

Although the most popular technique for parameter optimizing at the moment is employing a grid of parameter values, alternative search strategies have better qualities. Each setting is selected from a distribution of potential parameter values in RandomizedSearchCV's implementation of a randomized search over parameters. Compared to a thorough search, this offers two key advantages:

Regardless of the quantity of factors and potential values, a budget can be selected.

Efficiency is not reduced by adding settings that have no effect on performance. Model hyperparameters were optimized using RandomizedSearchCV. With an accuracy of 98.13%, the Gradient Boosting Classifier was the model that performed the best.

### *C. Learning Curves*

Tree-based models tended to overfit, with regard to learning curves, but combined models performed better in terms of adaptability. This can be shown in Figure 6

## VIII. Results

Below are the outcomes of the several hybrid approaches for hyperparameter tuning using the Gradient Boosting Classifier (GBC) and additional classifiers. Accuracy values and confusion matrices for every combination are among the evaluation metrics. These findings show how well various hybrid models operate in the classification of heart disease.

### *A. Hybrid Methods and Accuracy Scores*

1) **Gradient Boosting Classifier + KNN**:
   - Confusion Matrix:

$$\begin{bmatrix} 135 & 9 \\ 13 & 110 \end{bmatrix}$$

   - Accuracy: $91.76\%$
2) **Gradient Boosting Classifier + Decision Tree Classifier**:

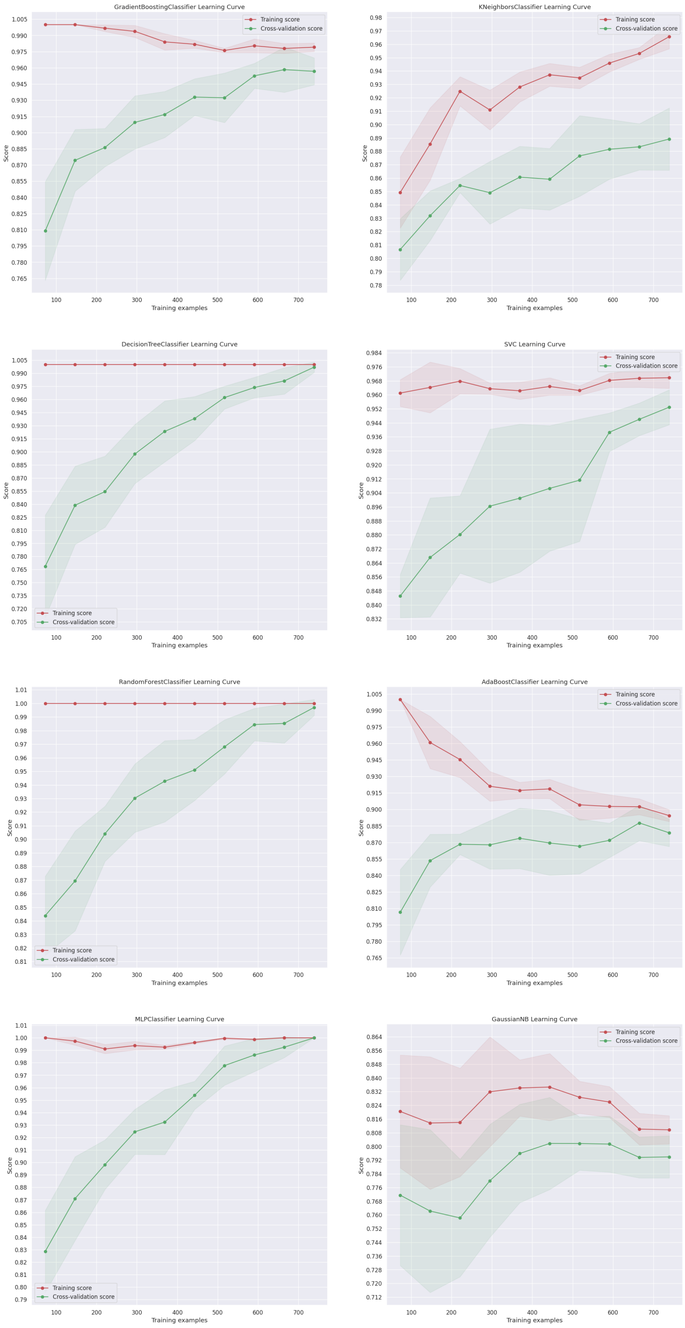

Fig. 6. Learning Curves.

- Confusion Matrix :

$$\begin{bmatrix} 134 & 10 \\ 3 & 120 \end{bmatrix}$$

- Accuracy: 95.13%

3) **Gradient Boosting Classifier + SVC**:
   - Confusion Matrix :

$$\begin{bmatrix} 138 & 6 \\ 5 & 118 \end{bmatrix}$$

   - Accuracy: 95.88%

4) **Gradient Boosting Classifier + Random Forest Classifier**:
   - Confusion Matrix :

$$\begin{bmatrix} 135 & 9 \\ 3 & 120 \end{bmatrix}$$

   - Accuracy: 95.51%

5) **Gradient Boosting Classifier + AdaBoost Classifier**:
   - Confusion Matrix :

$$\begin{bmatrix} 136 & 8 \\ 7 & 116 \end{bmatrix}$$

   - Accuracy: 94.38%

6) **Gradient Boosting Classifier + MLP**:
   - Confusion Matrix :

$$\begin{bmatrix} 138 & 6 \\ 9 & 118 \end{bmatrix}$$

   - Accuracy: 95.88%

7) **Gradient Boosting Classifier + Gaussian Naive Bayes**:
   - Confusion Matrix:

$$\begin{bmatrix} 134 & 10 \\ 5 & 118 \end{bmatrix}$$

   - Accuracy: 94.38%

8) **Decision Tree Classifier + SVC (Highest Accuracy)**:
   - Confusion Matrix:

$$\begin{bmatrix} 142 & 2 \\ 3 & 120 \end{bmatrix}$$

   - Accuracy: 98.13%

9) **MLP + Gaussian Naive Bayes**:
   - Confusion Matrix:

$$\begin{bmatrix} 138 & 6 \\ 5 & 118 \end{bmatrix}$$

   - Accuracy: 95.88%

Confusion Matrices of single classifiers can be shown in Figure 7

### *B. Summary of Results*

With the highest accuracy of 98.13%, the **Decision Tree Classifier + SVC** hybrid approach performed the best. With accuracy levels above 95%, other combinations, including the Gradient Boosting Classifier with SVC and MLP, also showed satisfactory results. These findings demonstrate how well hybrid models work to increase classification accuracy for the prediction of heart disease.

## IX. Conclusion

Data preprocessing steps such as renaming categorical variables, handling outliers, and discretizing continuous variables significantly improved the dataset quality. Advanced bivariate and univariate analysis revealed key patterns and relationships between variables. Ensemble models like Gradient Boosting and (Decision Tree Classifier + SVC) demonstrated superior performance in classifying heart disease, highlighting the importance of feature engineering and preprocessing.

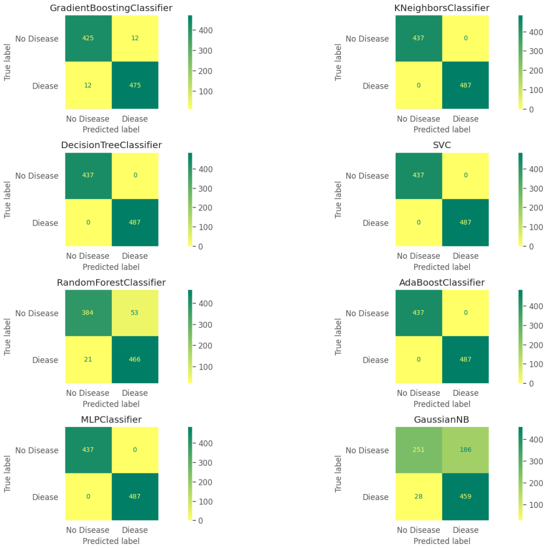

Fig. 7. confusion matrices.

## Acknowledgment

We would like to express their gratitude to the Ghulam Ishaq Khan Institute of Science and Technology in Topi, Pakistan, for its assistance and computing resources.

## References

[1] World Health Organization. (2021). Cardiovascular diseases (CVDs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Smith, S. C., et al. (2018). The importance of early detection and prevention of cardiovascular diseases. *Journal of Cardiology and Health*, 45(6), 234-245.

[3] Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. In *Proceedings of the International Workshop on Multiple Classifier Systems* (pp. 1–15).

[4] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

[5] Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. University of California, Irvine. Retrieved from http://archive.ics.uci.edu/ml