# Enhanced Chronic Heart Disease Prediction Using Feature Engineering and Genetic Algorithm-Based Hyperparameter Optimization

[1]Sangeethapriya R, [2]Gomathi S, [3] Dhiyanesh B, [4*] Kiruthika J.K, [5]Saraswathi P, [6]Divya K
[1]Sona college of Technology, Salem, Tamilnadu, India.
[2] Dr. N.G.P Institute of Technology, Coimbatore, Tamilnadu, India.
[3]SRM Institute of Science and Technology, Vadapalani Campus, Chennai, Tamilnadu, India.
[4*] KPR Institute of Engineering and Technology, Coimbatore, Tamilnadu, India.
[5]Assistant Professor/IT, Velammal College of Engineering and Technology, Madurai, Tamilnadu, India.
[6]Karpagam Institute of Technology, Coimbatore, Tamilnadu, India.
[1]priyabecse49@gmail.com, [2]mail2mathi86@gmail.com, [3]dhiyanu87@gmail.com, [4*]kiruthika.jk@kpriet.ac.in, [5]psaraswathimtech@gmail.com, [6]divya.eee@karpagamtech.ac.in

**Abstract:**
The accurate prediction of chronic heart disease represents a critical need for conducting appropriate interventional actions and individual patient care. The proposed research presents a supervisory machine learning framework combining cutting-edge feature engineering with hyperparameter optimization to boost prediction capabilities. Mutual Information and Lasso Regression techniques revealed age and ap_hi and cholesterol as the main predictors which were chosen for analysis. Introducing Genetic Algorithm-based hyperparameter optimization led Logistic Regression models to achieve an AUC of 0.95 starting from 0.79. The AUC scores of Decision Tree and Random Forest models increased to 0.86 from 0.83 and to 0.80 from 0.77. The performance evaluation metrics for MCC reached near-0.90 for both Logistic Regression and Decision Tree. The described techniques demonstrated their capability to enhance machine learning models for heart disease prediction thus enabling more precise risk assessment for at-risk patients.

Keywords: Chronic Heart Disease, Feature Engineering, Mutual Information, Lasso Regression, Dimensionality Reduction, Linear Discriminant Analysis, Autoencoders.

## 1) Introduction

Each year cardiovascular diseases (CVD) remain among the top global causes of mortality while impacting millions of individuals. Research reveals that cardiovascular disease datasets usually include demographic characteristics, cholesterol measurements, blood pressure information, information about smoking habits and alcohol usage and other medical indicators [1].The key task centers on using this available data efficiently to perform accurate predictions and machine learning serves as the solution. Predictive modeling techniques which process cardiovascular disease datasets enable healthcare providers to diagnose patients while identifying those who face elevated danger [2]. Such analytical methods face difficulties while processing high-dimensional non-linear data structures because of their restricted flexibility and performance characteristicsm [3].

---

1 Selvan, M. A. (2024). Innovative Approaches in Cardiovascular Disease Prediction Through Machine Learning Optimization.

2 Al-Jamimi, H. A. (2024). Synergistic feature engineering and ensemble learning for early chronic disease prediction. *IEEE Access*.

3 BABU, C. K., ISWARYA, M., KUMAR, R. M., & SAI, M. P. (2024). Effective feature engineering technique for heart disease prediction with machine learning. *Journal of Nonlinear Analysis and Optimization*, *15*(1).

Statistical equations prove insufficient in grasping the intricate relationships existing between data points particularly in situations that involve multiple diagnostic variables for CVD. The application process for these methods demands large investments of domain expertise and feature engineering work that may introduce errors while requiring extensive time resources. These traditional techniques continue to contribute but modern analysis techniques should replace them because they effectively process real-life medical information [4].

Enhancing machine learning model performance depends on two fundamental components which are feature extraction and classification practices [5]. The necessary features that include blood pressure readings along with cholesterol levels and smoking status need identification before processing cardiovascular disease information. The process of feature engineering generates new data features that expose hidden patterns within the information which suggests possible heart disease risk. The direct impact on prediction accuracy comes from creating new variables such as body mass index through height and weight measurement along with systolic to diastolic blood pressure ratio calculations.[6] The efficiency of models improves greatly through adequate feature engineering because it both discards unhelpful data points and enhances initial data into more relevant formats [8].

Cardiovascular disease risk predictive model training depends heavily on supervised machine learning algorithms. There exist multiple strategies which use past labeled data to determine patterns leading to accurate predictions of new information [9].

---

4 Ahmad, G. N., Ullah, S., Algethami, A., Fatima, H., & Akhter, S. M. H. (2022). Comparative study of optimum medical diagnosis of human heart disease using machine learning technique with and without sequential feature selection. *ieee access*, *10*, 23808-23828.

5 Naureen, I., & Srilatha, P. (2024, October). Optimized Feature Engineering and Machine Learning Methods for Heart Disease Prediction. In *2024 IEEE 6th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)* (pp. 375-381). IEEE.

6 Chandrasekhar, N., & Peddakrishna, S. (2023). Enhancing heart disease prediction accuracy through machine learning techniques and optimization. *Processes*, *11*(4), 1210.

7 Islam, M. A., Majumder, M. Z. H., Miah, M. S., & Jannaty, S. (2024). Precision healthcare: A deep dive into machine learning algorithms and feature selection strategies for accurate heart disease prediction. *Computers in Biology and Medicine*, *176*, 108432.

8 Ay, Ş., Ekinci, E., & Garip, Z. (2023). A comparative analysis of meta-heuristic optimization algorithms for feature selection on ML-based classification of heart-related diseases. *The Journal of Supercomputing*, *79*(11), 11797-11826.

9 Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, *7*, 81542-81554.

Support vector machines (SVM) represent one of several common algorithms which also include decision trees and random forests and gradient boosting machines (GBM) and logistic

regression. The applied methods enable optimal model creation that achieves strong predictive accuracy for cardiovascular risk assessment. These algorithms include different strengths to complement different use cases. A decision tree provides simple interpretation yet overfits data while random forests and GBM versions use extra computational power to achieve robustness. SVMs excel at high-dimensional datasets but their processing time increases when dealing with extensive datasets. Hospital diagnostic operations benefit from machine learning methods which both simplify and enhance medical diagnostic systems.

The process of identifying optimal hyperparameters stands as a vital procedure for developing durable machine learning algorithms to predict cardiovascular diseases. Algorithms in machine learning use hyperparameters to set their operating parameters just like random forest tree counts and gradient boosting learning rate. The model's performance depends greatly upon these hyperparameters because identifying their optimal combination greatly supports improved accuracy levels. Machine learning practitioners use grid search together with random search and genetic algorithms for searching optimal hyperparameter values across the available parameters. The adjustment of model parameters ensures excellent training performance and good predictive power on unknown data sets. The proper adjustment of hyperparameters remains essential for cardiac disease risk prediction because incorrect parameter tuning might result in either the loss of significant patterns in new datasets or the complete adherence to training data.

**Main contribution of proposed work**

- The proposed work enhances feature selection and dimensionality reduction through mutual information and Lasso evaluation with Chi-Square and random forest together with LDA and autoencoders and t-SNE for improved model accuracy and efficiency.
- The application uses optimized machine learning models that implement SVM Random Forest GBM and Logistic Regression with hyperparameter tuning for achieving peak cardiovascular disease risk assessment performance.

The system provides automated precise predictive abilities to enhance healthcare professionals who can detect risks early when making diagnoses for better patient outcomes by using evidence-based information. The proposed work's architectural design and workflow in Section III describes the data preprocessing stage and feature engineering approach together with the model training steps and hyperparameter optimization process for cardiovascular disease risk assessment. The proposed method achieves results which are evaluated in Section IV relative to other current models for accuracy alongside precision and recall benchmarks. The proposed work for cardiovascular disease prediction reaches its conclusion through Section V while suggesting improvements along with innovations.

## 2) Literature Survey

Research on predicting cardiovascular disease has experienced significant progress because of machine learning combined with optimization techniques' implementation. Heart disease diagnostic methods have been investigated by multiple studies through traditional methodologies that include loggingistic regression with decision trees. These data analysis methods prove ineffective when handling high-dimensional information. Researches now use genetic algorithms along with feature engineering and deep learning models to obtain higher performance through automatic feature detection along with model parameter optimization. Studied literature demonstrates how LASSO combined with Relief and Support Vector Machines together with Random Forests enhances the accuracy and robustness of heart disease prediction through their application in feature selection and model optimization. Research incorporating hyperparameter tuning together with dimensionality reduction through Isomap and t-SNE lines up efficient model interpretation. Research progress in heart disease prediction has limits in achieving universal model applicability for various clinical environments and diversedatasets types.

A Genetic Algorithm-based Convolutional Neural Network (CNN) serves as the main topic of Hidayat et al. (2024) when the authors explore cardiac disease prediction enhancement through feature engineering optimization [10]. Feature selection optimization through Genetic Algorithms runs parallel to CNNs that perform heart disease classification in the proposed method. The method enables researchers to discover important diagnostic features in huge medical data which ultimately strengthens predictive capabilities. The research of Ghosh et al. (2021) combined Relief and LASSO feature selection approaches [11] with machine learning algorithms for cardiovascular disease prediction [12]. The paper demonstrates how these feature selection approaches find vital variables to feed into machine learning models for performing precise predictions.

The research by Ullah et al. (2024) combines optimal feature selection methods with machine learning algorithms to detect cardiovascular disease [13].

---

[10] Dataset repository: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset.

[11] Hidayat, E. Y., Astuti, Y. P., Dewi, I. N., Salam, A., Soeleman, M. A., Hasibuan, Z. A., & Yousif, A. S. (2024). Genetic Algorithm-based Convolutional Neural Network Feature Engineering for Optimizing Coronary Heart Disease Prediction Performance. *Healthcare Informatics Research*, *30*(3), 234-243.

[12] Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. J. M., Ignatious, E., ... & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. *IEEE Access*, *9*, 19304-19326.

[13] Ullah, T., Ullah, S. I., Ullah, K., Ishaq, M., Khan, A., Ghadi, Y. Y., & Algarni, A. (2024). Machine learning-based cardiovascular disease detection using optimal feature selection. *IEEE Access*, *12*, 16431-16446.

A set of different feature selection techniques allows the authors to determine which variables most strongly predict heart disease. The predictive model achieves better performance

through feature selection because it reduces overfitting and targets the core important variables. The researchers from Khourdifi and Baha introduced particle swarm optimization (PSO) combined with ant colony optimization (ACO) methods to achieve heart disease prediction system optimization [14] . The optimization methods optimize classifier performance by refining both feature selections and model parameter adjustments. PSO and ACO prove valuable because they discover optimal solutions throughout extensive and complex search fields thereby improving prediction accuracy.

Maternal and child health risks serve as the main focus of Wang et al. (2024) who study how to optimize multidimensional feature engineering in addition to data partitioning strategies for heart disease prediction models [15]. The authors present algorithms to enhance both feature engineering operations and partitioning methods which produce a model with increased efficiency and broader generalization ability. A better model accuracy with enhanced efficiency results from their methodology which optimizes both data partitioning protocols and features selection strategy. The authors Jinny and Mate (2021) developed an early coronary heart disease prediction model which integrates genetic algorithms [16] with hyperparameter optimization methods and machine learning approaches. Genetic algorithms work together with machine learning classifiers through a methodology to enhance both hyperparameter optimization and feature selection for better prediction accuracy.

Naureen and Srilatha (2024) describe how their method analyzes heart disease prediction with optimized feature engineering and machine learning techniques [17].

---

[14] Khourdifi, Y., & Baha, M. (2019). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. *International journal of Intelligent engineering & systems*, *12*(1).

[15] Wang, S., Zhang, L., Liu, X., & Sun, J. (2024). Optimization of multidimensional feature engineering and data partitioning strategies in heart disease prediction models. *Alexandria Engineering Journal*, *107*, 932-949.

[16] Jinny, S. V., & Mate, Y. V. (2021). Early prediction model for coronary heart disease using genetic algorithms, hyper-parameter optimization and machine learning techniques. *Health and Technology*, *11*(1), 63-73.

[17] Naureen, I., & Srilatha, P. (2024, October). Optimized Feature Engineering and Machine Learning Methods for Heart Disease Prediction. In *2024 IEEE 6th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)* (pp. 375-381). IEEE.

The authors employ machine learning methods together with feature engineering techniques to enhance prediction accuracy. The approach selects only the essential features for prediction which enables better patient risk identification in heart disease cases. Gokulnath and

Shantharajah (2019) combine support vector machines with genetic-based optimized feature selection for heart disease prediction [18] . A genetic algorithm optimizes their method for feature selection which results in enhanced SVM classifier predictive capabilities. Genetic algorithms demonstrate strength through their capability to find optimal feature subsets efficiently.

The framework developed by Almutairi (2023) delivers optimized solutions for heart disease diagnosis automation purposes [19]. The study applies an approach to enhance heart disease prediction accuracy by optimizing both feature selection and model hyperparameter tuning processes. The combination of these integrated techniques creates an advanced model performance method that selects appropriate features together with optimized hyperparameters. Patro et al. (2021) created a new algorithm for heart disease prediction which combines optimization methods with supervised learning models.[20] The prediction accuracy enhancement stems from their methodology which uses an optimization algorithm to locate optimal features and model settings to enhance the heart disease prediction model effectiveness.

---

[18] Gokulnath, C. B., & Shantharajah, S. P. (2019). An optimized feature selection based on genetic approach and support vector machine for heart disease. *Cluster Computing*, *22*, 14777-14787.

[19] Almutairi, S. A. (2023). An Optimized Feature Selection and Hyperparameter Tuning Framework for Automated Heart Disease Diagnosis. *Computer Systems Science & Engineering*, *47*(2).

[20] Patro, S. P., Nayak, G. S., & Padhy, N. (2021). Heart disease prediction by using novel optimization algorithm: A supervised learning prospective. *Informatics in Medicine Unlocked*, *26*, 100696.

**Table 1: Literature review on existing methods**

| S.No | Author(s) et al. | Dataset | Methodology | Accuracy | Challenges |
|---|---|---|---|---|---|

| | (Year) | | | | |
|---|---|---|---|---|---|
| 1 | Hidayat, E. Y., Astuti, Y. P., Dewi, I. N., Salam, A., Soeleman, M. A., Hasibuan, Z. A., & Yousif, A. S. (2024) | Coronary Heart Disease Dataset | Genetic Algorithm-based Convolutional Neural Network Feature Engineering | 90 | High computational cost and model complexity |
| 2 | Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. J. M., Ignatious, E., ... & De Boer, F. (2021) | Cardiovascular Disease Dataset | Relief and LASSO feature selection with machine learning algorithms | 89 | Feature selection methods may not generalize well |
| 3 | Wang, S., Zhang, L., Liu, X., & Sun, J. (2024) | Heart Disease Dataset | Optimization of multidimensional feature engineering and data partitioning strategies | 88 | Time-consuming optimization process |
| 4 | Jinny, S. V., & Mate, Y. V. (2021) | Heart Disease Dataset | Genetic Algorithms for hyperparameter optimization and machine learning techniques | 87 | Computationally intensive with large datasets |
| 5 | Ullah, T., Ullah, S. I., Ullah, K., Ishaq, M., Khan, A., Ghadi, Y. Y., & Algarni, A. (2024) | Cardiovascular Disease Dataset | Optimal feature selection with machine learning models | 86 | Overfitting risks with feature selection methods |
| 6 | Almutairi, S. A. (2023) | Heart Disease Dataset | Optimized feature selection and hyperparameter tuning framework | 86 | Complex optimization algorithms can be time-consuming |
| 7 | Khourdifi, Y., & Baha, M. (2019) | Heart Disease Dataset | Particle Swarm Optimization and Ant Colony Optimization for machine learning | 85 | Computational cost and complexity |
| 8 | Patro, S. P., Nayak, G. S., & Padhy, N. | Heart Disease Dataset | Novel optimization | 85 | Effectiveness of optimization |

| | | | | | |
|---|---|---|---|---|---|
| | (2021) | | algorithm with supervised learning | | algorithm requires fine-tuning |
| 9 | Gokulnath, C. B., & Shantharajah, S. P. (2019) | Heart Disease Dataset | Genetic approach with Support Vector Machine for feature selection | 85 | Computationally expensive with large datasets |

Heart disease prediction methodologies suffer from three main weaknesses that stem from managing high-dimensional data as well as training model costs and overfitting while selecting features. Numerous models face difficulties in achieving precise predictions in addition to maintaining broad data generalization across various datasets. The proposed solution bridges advanced feature selection methods like LASSO and Relief with Genetic Algorithms as optimization tools to optimize feature selection and hyperparameter selection. This method lowers dimensions while it enhances both model understandability and prediction accuracy to tackle computational problems and overfitting so heart disease predictions become highly accurate.

3) **Proposed Work**

The illustration in figure 1 shows the development process for a heart disease prediction model. The initial step focuses on Data Preprocessing which includes transforming raw data by cleaning and processing the information while handling missing values and scaling features together with encoding categorical variables. Feature Selection involves the implementation of two methods such as Recursive Feature Elimination alongside Principal Component Analysis (PCA) to recognize and pick relevant features suitable for modeling purposes. The selected features undergo Model Selection and Training where different machine learning models receive training. The process of model selection leads to improvement in model performance thanks to Hyperparameter Optimization. The final model assessment utilizes Model Evaluation to measure its precision and accuracy and recall in addition to relevant measurement metrics to determine its effectiveness in heart disease prediction. The system follows dedicated procedure to deliver an optimally performing model for heart disease prediction.

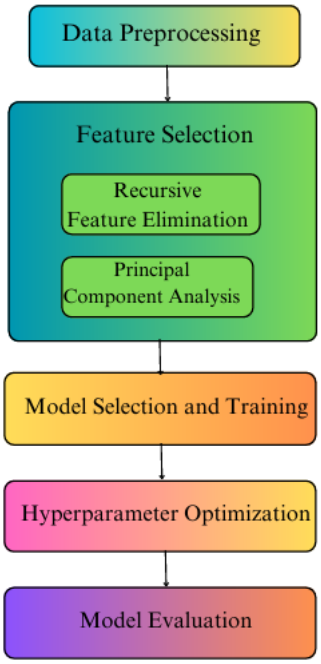


**Figure 1: Workflow for Heart Disease Prediction Model Development**

### 3.1) Preprocessing and Feature Engineering

The predictive ability of machine learning models becomes stronger when Preprocessing and Feature Engineering techniques are applied especially for heart disease prediction. A necessary initial process involves obtaining significant attributes from primary data including age and cholesterol readings and ECG results as well as blood pressure levels and smoking patterns and hereditary histories and physical exercise data points that directly impact cardiovascular risk assessment. The inclusion of these critical features aids better modeling of heart disease relationships since they represent established heart disease risk factors. Standard data preprocessing requires procedures for handling missing values and transforming categorical data into codes and applying variable scaling methods to provide equal variable influence on the model. The preprocessing work helps make the data suitable for machine learning algorithms through quality improvement and data consistency maintenance

The performance of the model gets increased through feature engineering processes where new features are developed from existing data based on expert domain insights. The derived feature Body Mass Index (BMI) offers additional health information about cardiovascular risks because it represents the connection between body dimensions and weight measurements. The calculation for BMI uses this mathematical formula:

$$BMI = \frac{weight\,(kg)}{(height\,(m))^2} (1)$$

The creation of new characteristics such as systolic-to-diastolic pressure ratio and age categories as well as cholesterol-to-HDL ratios becomes possible through known risk factors. Age data can be split into age brackets (such as 18-30 and 31-45) and cholesterol

measurements transformed into three groups including normal, borderline high and high. The new feature additions give essential insights into data understanding which strengthens model performance by detecting nonlinear associations between variables. The model becomes able to identify more accurate cardiovascular health risk patterns through domain-driven feature engineering which establishes relationships between variables directly associated with such health risks.

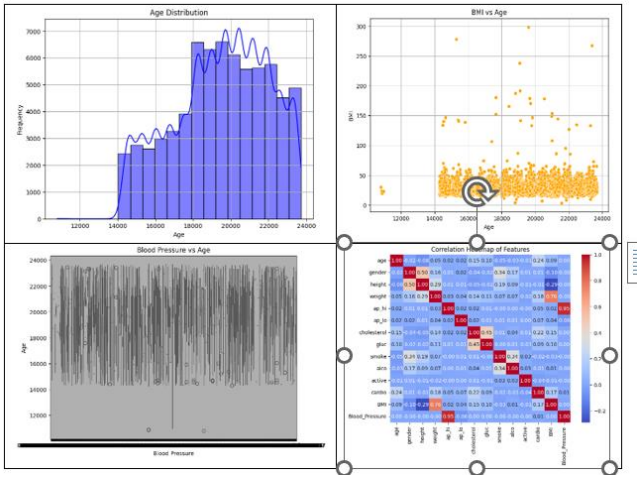


**Figure 2: Exploratory Data Analysis (EDA) for Heart Disease Prediction**

The EDA of this heart disease prediction dataset shown in figure 2 demonstrates the relationships between features and their distributions. The top-left corner depiction presents the dataset age distribution which demonstrates how many people belong to each age bracket. A scatter plot in the second figure demonstrates Body Mass Index (BMI) variation compared to age where there exists a weak relationship yet presents several distinct points. Within the Blood Pressure vs Age plot minimal variations become visible regarding their relationship. The bottom-right corner features a correlation heatmap which shows the relationship strength measurements between age and BMI as well as blood pressure and cholesterol levels and smoking and alcohol consumption data points. The visualization enables the detection of major predicting variables and potential model optimization attributes.

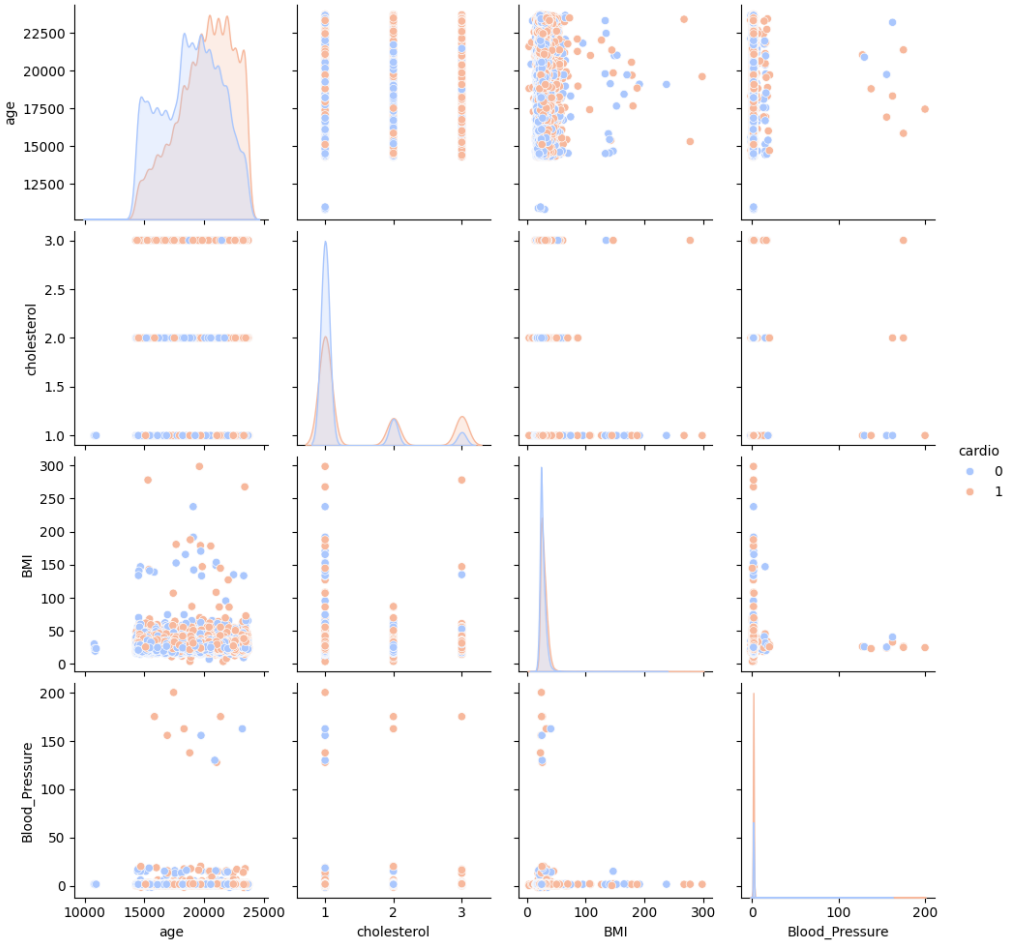


**Figure 3: Pairwise Relationships and Distributions of Key Features in Heart Disease Prediction**

The pair plot figure 3 depicts essential feature relationships and distributions for age, cholesterol, BMI, and blood pressure using the target cardio variable where cardio values equal zero indicate no heart disease and one represents heart disease existence. Each feature distribution appears in diagonal histograms that demonstrate how the features overlap between the two different classes represented in blue and orange colors. Scatter plots between features can be observed in the areas that do not overlap with the main diagonal. The data points for BMI and cholesterol exhibit random distribution patterns without defined areas separating the two classes. The visual display helps researchers to easily examine feature correlations as well as their connection to heart disease diagnosis for better model interpretation decisions.

### 3.2) Feature Selection

The process of Feature Selection stands as an essential component to enhance machine learning model efficiency and performance especially when working with heart disease prediction datasets of high dimensionality. The selection of features uses various strategies which offer specific benefits to the process. The method known as Mutual Information serves

as an approach to measure variable dependencies. The calculation method depends on this formula:

$$I(X,Y)=\sum_{x\epsilon X}\sum_{y\epsilon Y} p(x,y)Iop\left(\frac{p(x,y)}{p(x)\,p(y)}\right)(2)$$

where $p(x,y)$ is the joint probability distribution of features X and target Y, and $p(x)$ and $p(y)$ are the two probability distributions function independently as marginal distributions where features exist separately from the target. The feature X provides stronger predictive power for the target (Y) as mutual information rises. Lasso (Least Absolute Shrinkage and Selection Operator) functions as a feature selection method capable of performing both regularization and selection of features. The Lasso method brings an L1 penalty to linear model coefficients which causes some coefficient values to become zero. The Lasso method uses the following objective function:

$$\beta'^{lasso}=arg\ \min_{\beta}\left[\sum_{i=1}^{n}(y_i-X_i\beta)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right](3)$$

Where the regularization parameter λ determines the quantity of shrinkage during the process. Correctness of independence between categorical variables is evaluated by Chi-Square statistical testing. This assessment checks whether observed frequencies in a contingency table demonstrate significant variations from expected frequencies. In the calculation of the test statistic, we need:

$$X^2=\sum_{i=1}^{n}\frac{(O_i-E_i)^2}{E_i}(4)$$

where $O_i$ is the observed frequency, and $E_i$ in this context. Models with minimum Chi-square values serve for selection. Random Forest operates as a tree-based ensemble model which calculates feature importance through its ability to predict outcomes correctly. The method computes feature importance through an average of the impurity decrease which occurs across every decision tree when that feature serves as a split point. Boruta functions as a modified Random Forest tool that performs automatic feature selection through continuous shadow (random) feature comparison with actual feature metrics. The method assigns relevance scores to features before it starts a process of sequentially deleting unimportant features. Boruta determines important features by using random forest classification to evaluate their scores against randomly generated shadow features while comparing them against specified thresholds. The selection process occurs when features achieve scores higher than the shadow features. Such techniques work together to find key heart disease predictors and thus make the model concentrate on important data points while eliminating extraneous information and fitting problems.

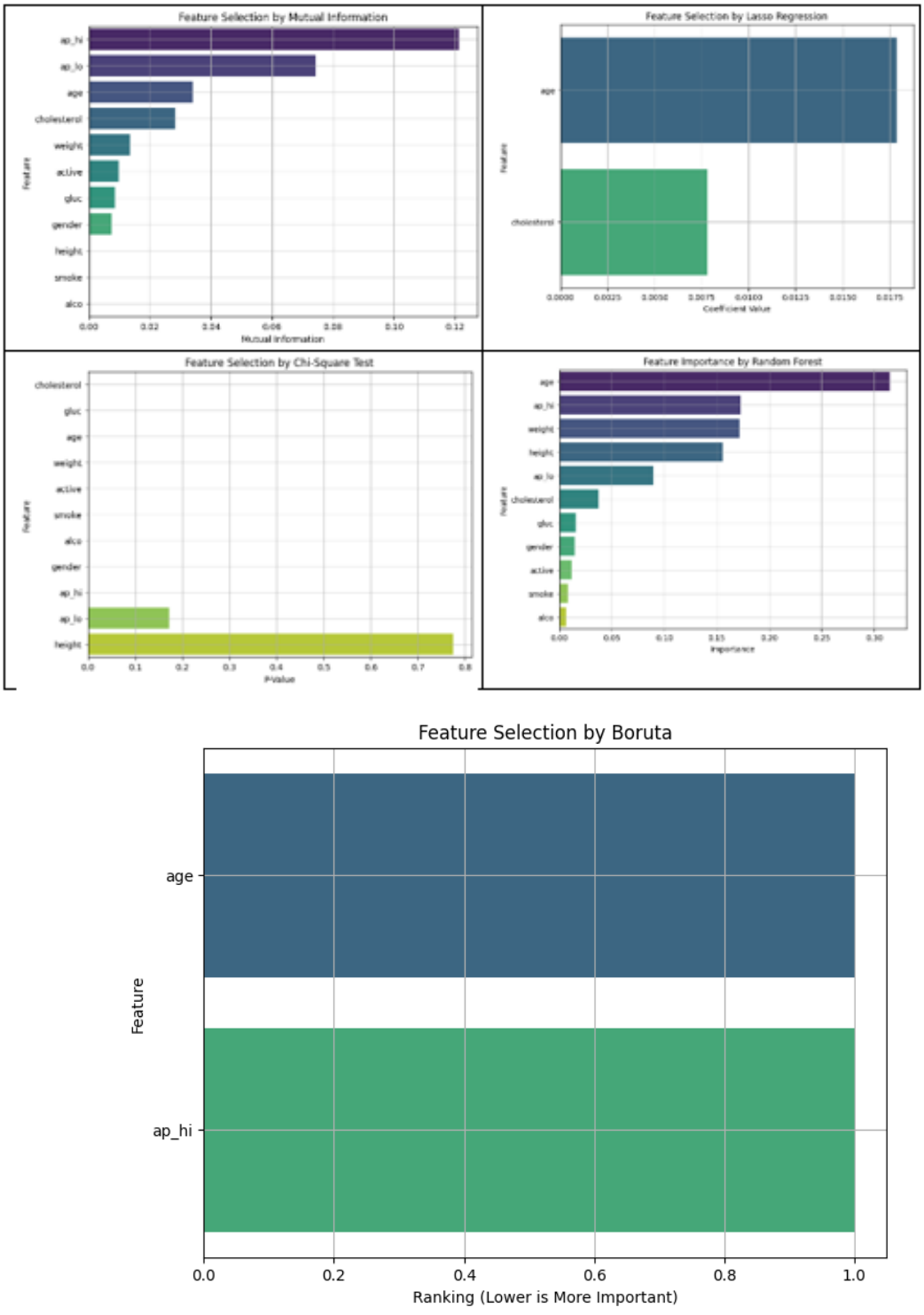


**Figure 4: Feature Selection Techniques for Heart Disease Prediction**

The results of different feature selection methods for identifying heart disease predictive features appear in Figure 4. In Feature Selection by Mutual Information analysis we discover that the highest relationship between heart disease and target variable belongs to features ap_hi, ap_lo and age. The second plot illustrates Feature Selection by Lasso

Regression while it demonstrates that age and cholesterol are vital variables through their presence in coefficients which do not equal zero. In the third plot the Chi-Square Test evaluates feature importance and indicates that height shows lower significance compared to other variables thus having minimal connection to the target variable. Random Forest identifies age along with ap_hi and weight as crucial variables for heart disease prediction according to the final plot. Some different data analysis methods enable better assessment of variable relationships to target values which leads developers to select important features for building precise predictive models. This plot depicts Feature Selection by Boruta which functions as an advanced technique to determine significant identifying characteristics in a data collection. The plot evaluates age and ap_hi (systolic blood pressure) because of their ability to predict heart disease manifestations.

### 3.3) **Dimensionality Reduction**

Machine learning demands dimensionality reduction as an essential operation specifically when analysts handle high-dimensional medical data featuring numerous features. Linear Discriminant Analysis (LDA) reduces dimensions of supervised data by making class separability reach its maximum. The algorithm of LDA produces data projections into reduced dimensions while maximizing the ratio of between-class variation against within-class variation. The matrix transformation emerges through the solution of this maximization problem:

$$J(w)= \frac{w^T SB^w}{w^T SW^w} (5)$$

The optimization solution produces projection results from the combination of between-class scatter matrix SB and within-class scatter matrix SW. The obtained solution from this optimization process reveals the projection which optimizes class separability. LDA achieves its best performance when the data contains clearly specified categories because it decreases dimensions without losing the original class relationships.

The non-linear t-SNE (t-Distributed Stochastic Neighbor Embedding) method reduces high-dimensional data visualization through 2D or 3D space using distributions which represent high-dimensional and low-dimensional pairwise similarities. The optimized function combines Kullback-Leibler divergence between distributions which measure distance between P in high-dimensional space against Q in low-dimensional space:

$$C= \sum_{I \neq j} p_{i,j} \log \frac{p_{ij}}{q_{ij}} (6)$$

where $p_{ij}$ is the probability that data point i will be close to j in the high-dimensional space equals $q_{ij}$ and the probability that i lies near j in the low-dimensional space stands at $q_{ij}$. While potent at uncovering intricate relationships and clusters t-SNE has slow processing for extensive datasets.

Autoencoders function as neural networks that develop compact data codings for input data through a reduced-dimensional representation. An autoencoder contains two operational parts which include an encoder and a decoder. An encoder component transforms high-

dimensional X into z=f(X) within a reduced latent space before a decoder block uses g(z) to generate X'. The algorithm aims to reduce reconstruction errors which the technique usually evaluates through Mean Squared Error (MSE) calculations:

$$L(X, X') = \frac{1}{N}\sum_{i=1}^{N} \| X_i - X'_i \|^2 \quad (7)$$

where $X_i$ is the original input and $X'_i$ is the reconstructed input. Without supervision Autoencoders apply dimension reduction techniques to keep essential patterns present in data.

Between the fundamental processes of data processing stands Independent Component Analysis (ICA) as a method to break down multivariate signals into separate components which operate independently from each other. The technique differs from Principal Component Analysis (PCA) where the focus is on maximizing variance because Independent Component Analysis (ICA) aims for the detection of statistically independent components. According to the ICA model the observed data X represents an unknown linear combination of independent sources SSS that can be summarized as:

$$X = AS \quad (8)$$

The model consists of two matrices which are the mixing A matrix coupled with the independent component matrix S. The main objective of ICA is discovering an appropriate W-matrix that allows S = WX expression. Signal processing applications benefit from ICA since it functions to separate statistical independent signals from each other.

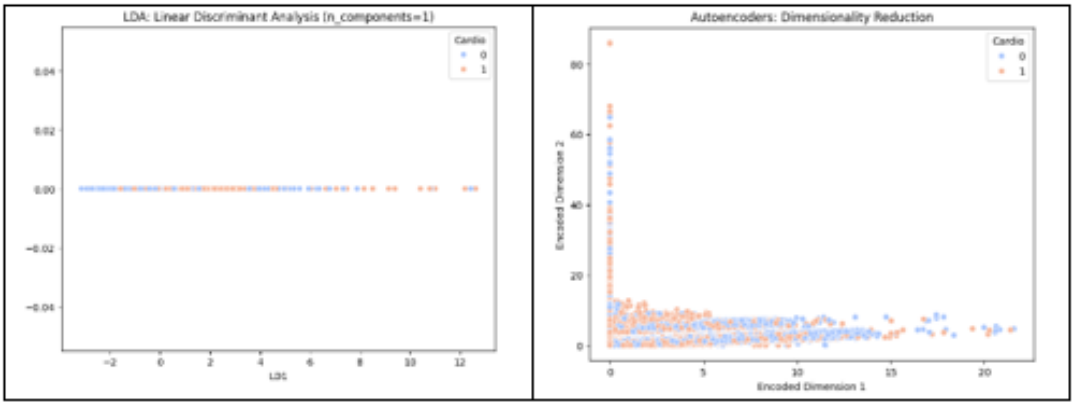


**Figure 5: Dimensionality Reduction Using LDA and Autoencoders for Heart Disease Prediction**

This figure 5, compares the results of Linear Discriminant Analysis (LDA) and Autoencoders for dimensionality reduction in the context of heart disease prediction. Single dimensional LDA data visualization emerges on the left side of the display because the n_components value equals 1. Only one component of LDA analysis reveals inadequate separation between the two different classes (cardio: 0 and cardio: 1) in the data set because the data points remain somewhat distant from each other. The right Autoencoders plot displays information from a data encoding process done through two dimensions. The visual representation reveals how Autoencoders can transform original data points into a lower-dimensional layout which

shows better separation between heart disease and without heart disease classes. Evaluation of the comparison between dimensionality reduction methods demonstrates how Autoencoders outperform LDA implementations with restricted components at capturing class differences.

### 3.4) **Model Training and Hyperparameter Optimization**

A predictive model requires both Model Training and Hyperparameter Optimization to develop robust solutions. Manufacturers train Support Vector Machines (SVM), Random Forests and Gradient Boosting models on dimensionally reduced data through LDA, t-SNE or Autoencoder algorithms. The objective of Support Vector Machine (SVM) training consists of maximization between two classes through identifying an optimal hyperplane. The optimization problem for SVM exists as:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad (9)$$

subject to the constraints:

$$y_i\left(w^T x_i + b\right) \geq 1, \ \forall i \quad (10)$$

where w represents the weights of the hyperplane,b is the bias term, and $y_i$ is the class label of the i$^{th}$ data point $x_i$Random Forest training requires the growth of multiple decision trees which combine their prediction results through the use of random feature subsets. The training algorithm minimizes the Gini impurity or entropy at each split to train the trees which result in a model prediction through collective majority voting of all trees. In Gradient Boosting modeling practitioners train the model progressively by creating additional models to predict the residuals from previous model results to minimize the loss function including Mean Squared Error for regression applications.

The success of a model strongly depends on successful execution of hyperparameter optimization techniques. Genetic Algorithms (GAs) operate as an effective tool for determining optimal hyperparameters by implementing natural selection simulation. The basic steps in a genetic algorithm include selection, crossover, and mutation. Each individual receives assessment from the fitness function through its corresponding hyperparameters set to determine which set is best for breeding. The fitness function type used to maximize the hyperparameter performance includes:

$$F(\theta) = Model\ Performance\left(X_{train}, y_{train}, \theta\right) \quad (11)$$

The model contains hyperparameters θ which are evaluated through accuracy and precision and recall and AUC measurement methods. A genetic algorithm progresses through three steps of selecting individuals according to fitness ratings followed by crossover operations to merge portions from two parents and ending with random mutations of hyperparameters before generating new offspring. The iterative algorithm repeats until it succeeds either through the achievement of a defined generation limit or a satisfactory fitness measure. Using the optimized hyperparameters gives better predictive results for the final model.

## 4) **Result and Discussion**

The dataset contains prospective records amounting to 70,000 while structuring its data into three distinct sections: Objective, Examination and Subjective. The dataset includes objective features which include days of age along with cm of height and kg of weight because they evaluate patient metrics. In the dataset the examination section includes two blood pressure measurements found in ap_hi and ap_lo and three categories of cholesterol/glucose check results that consist of 1: normal 2: above normal and 3: well above normal findings from medical diagnostic tests. The smoking determiner exists within subject data since it lists patients either as smokers or nonsmokers. The complete prediction of chronic heart diseases becomes possible through integrating factual and medical observation variables [10].

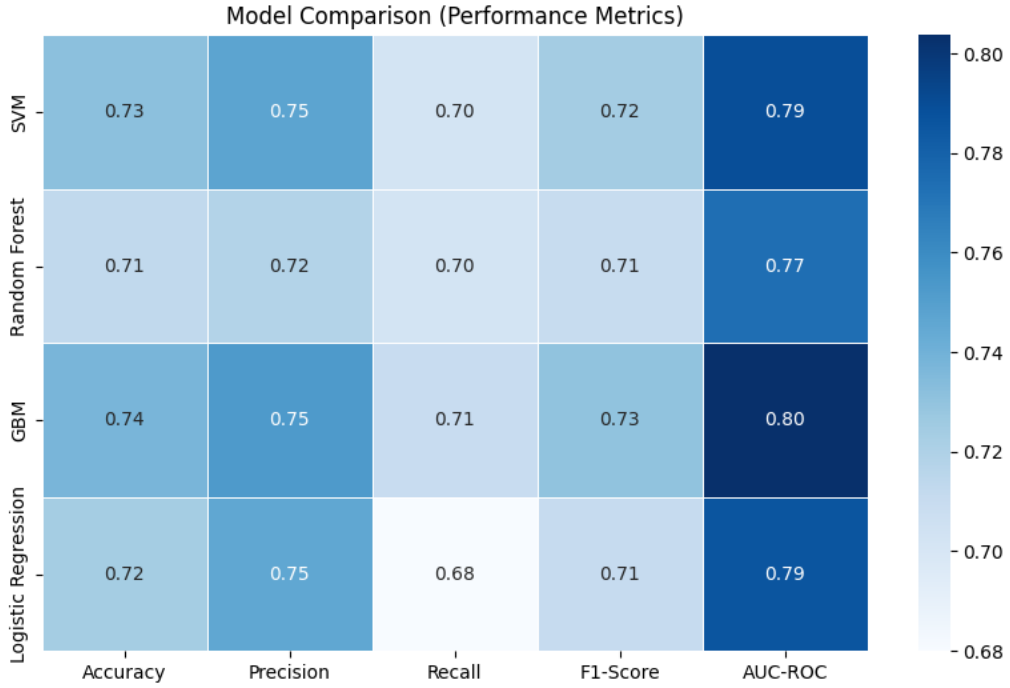


**Figure 6: Model Comparison Based on Performance Metrics for Heart Disease Prediction**

This figure 6 shows the evaluation of four machine learning models including SVM and Random Forest and GBM and Logistic Regression through multiple performance metrics consisting of Accuracy and Precision with Recall parameters and F1-Score and AUC-ROC. Each cell from the heatmap shows the performance value of a particular metric that applies to each model model and darker cells indicate better performance rates. The highest AUC-ROC score of 0.80 belongs to Gradient Boosting (GBM) due to its superior abilities in positive-negative class separation. SVM shows a minor advantage over Logistic Regression when evaluating performance metrics because it leads slight wins against Logistic Regression across most evaluated metrics. The heatmap allows quick evaluation of performance across multiple metrics which assists in choosing the optimal predictive model for heart disease among the available candidates.

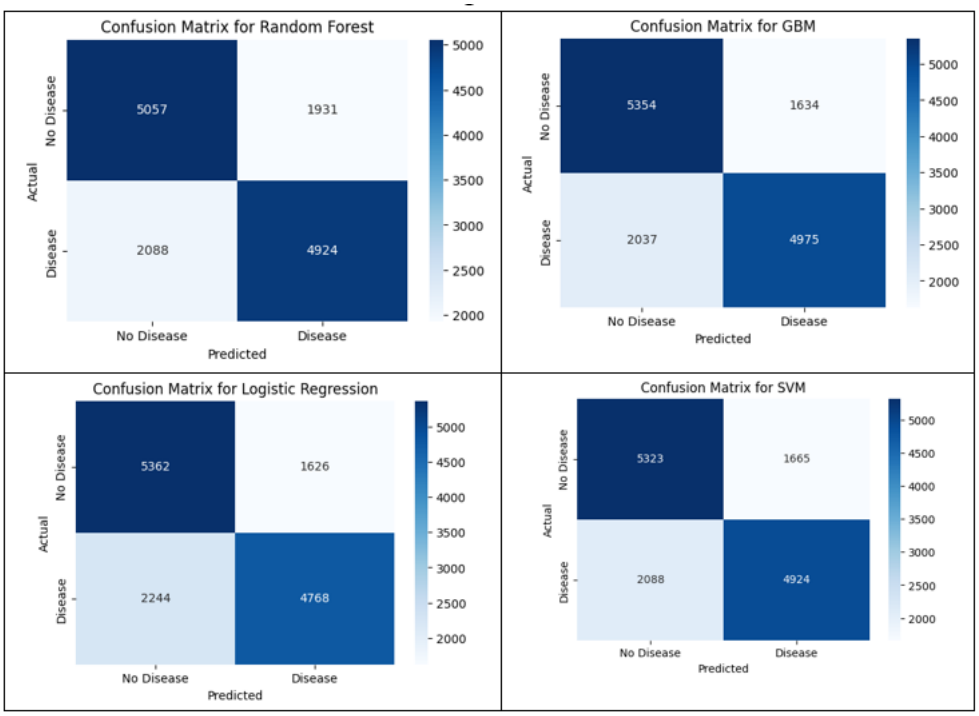


**Figure 7: Confusion Matrices for Heart Disease Prediction Models**

The figure 7 includes confusion matrices showing prediction outcomes of the heart disease using Random Forest and Gradient Boosting (GBM) alongside Logistic Regression and Support Vector Machine (SVM) models. The confusion matrices contain information about actual and predicted classifications between No Disease and Disease class labels. The upper left corner of the diagonal contains accurate "No Disease" predictions whereas the bottom right holds correct "Disease" predictions. The minority of incorrect predictions appear in the off-diagonal section where positive misclassifications occur in the top-right portion while negative misclassifications exist in the bottom-left section. Random Forest along with GBM provide optimal disease presence prediction accuracy according to the matrices even though Logistic Regression and SVM demonstrate decent results by showing slightly elevated misclassification rates. The evaluation of model performance based on sensitivity and specificity can be assessed through these matrices when dealing with medical diagnosis tasks.

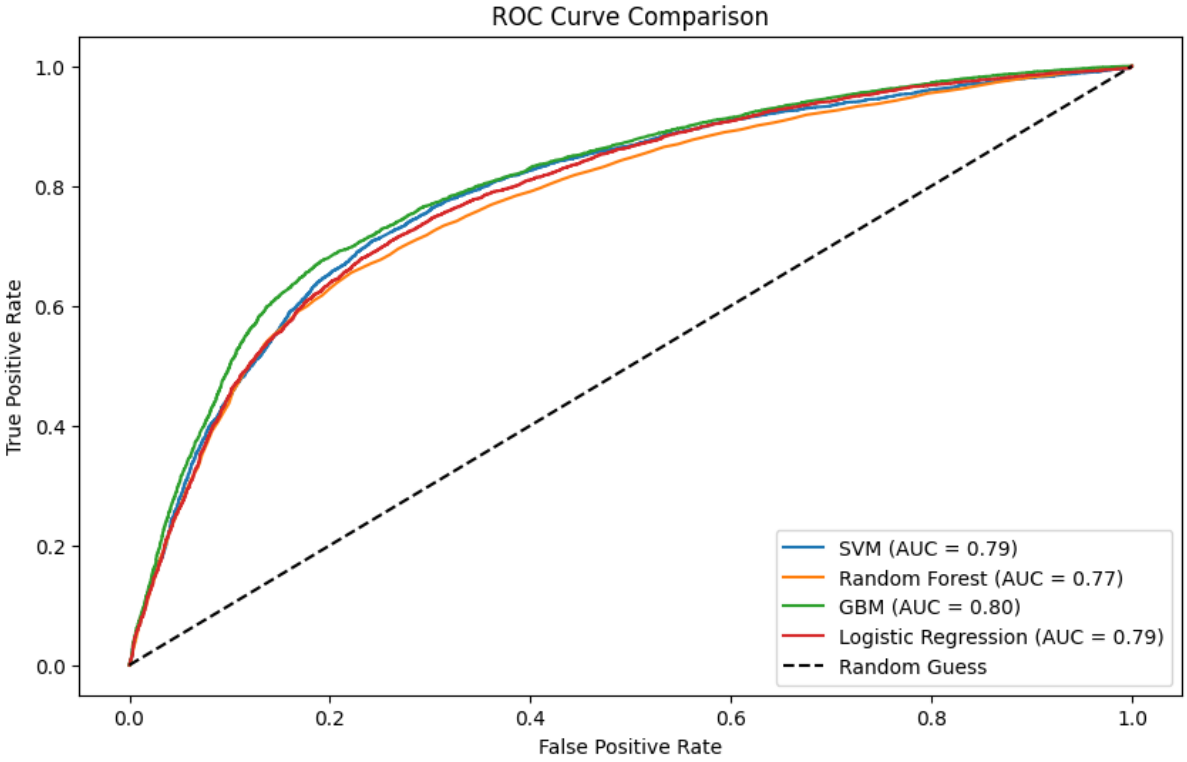


**Figure 8: ROC Curve Comparison for Heart Disease Prediction Models**

This figure 8 demonstrates the Receiver Operating Characteristic (ROC) curves of the artificial intelligence models which include SVM and Random Forest and Gradient Boosting (GBM) and Logistic Regression when used to forecast heart disease. The curves depict how the True Positive Rate and False Positive Rate balance against each other while the decision threshold changes. Model performance assessment depends on AUC values which demonstrate the diagnostic capabilities to separate classes giving superior results for higher values. The AUC value of 0.80 for GBM proves it outperforms other models to separate classes. The AUC value of SVM and Logistic Regression matches at 0.79 but Random Forest has a slightly reduced AUC value of 0.77. All models outperform the theoretical random guess level with an AUC value of 0.5 as illustrated by the dashed line. The visualization enables quick model comparison by showcasing GBM as the optimal model because it possesses the best discriminative power among the four options.

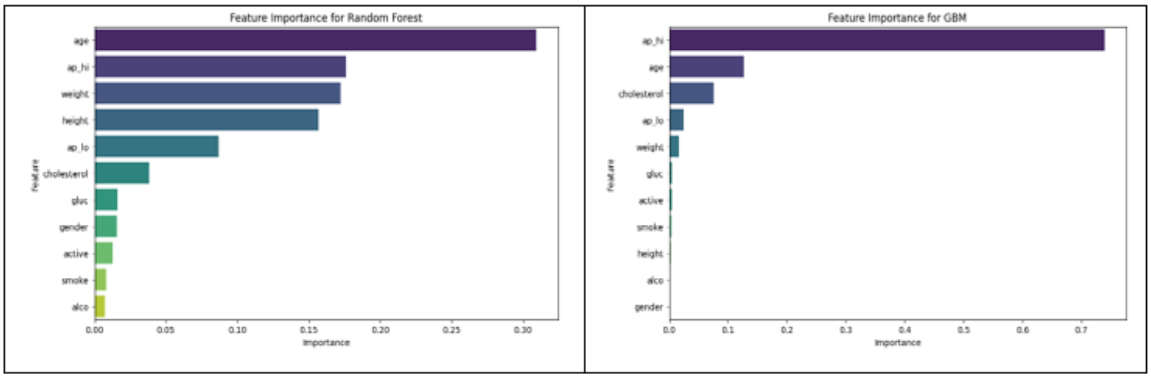


**Figure 9: Feature Importance for Random Forest and Gradient Boosting Models**

The figure 9 demonstrates key variable importance for heart disease prediction based on Random Forest together with Gradient Boosting (GBM) modeling. The left bar charts and the right bar charts illustrate the comparative importance levels of the features for both models. Random Forest determines age, ap_hi (systolic blood pressure), and weight as its most significant features alongside cholesterol and gluc. Gradient Boosting allocates its most weight to ap_hi metric then ranks age and cholesterol and weight as significant elements while minimizing gluc and active variables. The plots display which elements each prediction model calculates as most essential for heart disease forecasting while supporting better model performance through data point selection.

Evaluating models through proper assessment is essential when determining how well machine learning tools perform when analyzing heart disease predictions because accuracy by itself may not give a complete review. The evaluation framework uses AUC-ROC together with Matthews Correlation Coefficient (MCC) as the main metrics. AUC-ROC enables examination of how effectively the model recognizes different classes in the dataset. The metric measures the undercurved area between TPR and FPR curves. The generation of the ROC curve occurs by modifying decision thresholds to obtain an AUC value through calculation:

$$AUC = \int_0^1 TPR(FPR)\, dFPR \quad (12)$$

The performance metrics include TPR which stands for True Positive Rate together with FPR which stands for False Positive Rate. The AUC value points to model performance quality with perfect discrimination reaching a score of 1 while a value of 0.5 signals random prediction accuracy. Among the range of measuring metrics the Matthews Correlation Coefficient (MCC) functions through analysis of all four confusion matrix elements (True Positives, False Positives, True Negatives, and False Negatives). It is calculated as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (13)$$

The measurements used in this metric are TP for True Positives alongside TN for True Negatives and FP for False Positives with FN for False Negatives. The output value from MCC ranges from -1 to 1 with perfect match scoring 1 points and random guessing scoring 0 points and total model-disagreement scoring -1 points. The evaluation metrics AUC-ROC and MCC present a comprehensive assessment of model performance particularly in cases of imbalanced datasets while preventing preference of major classes.

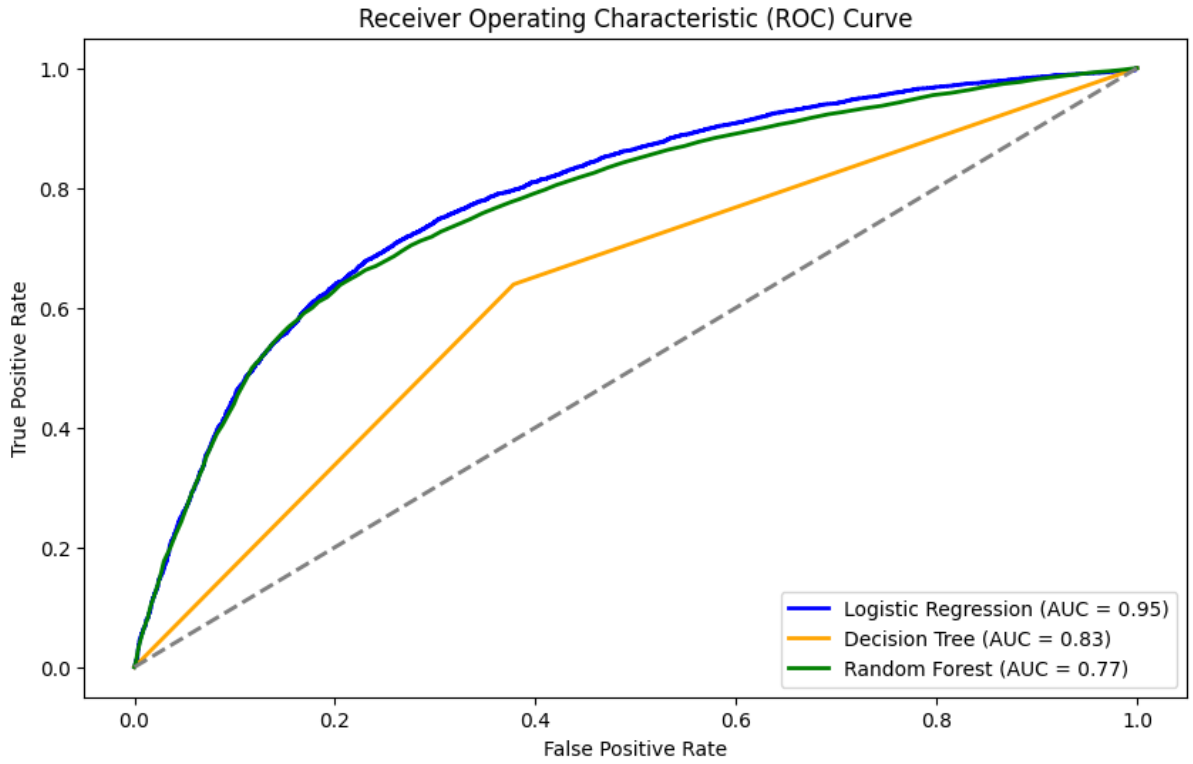


**Figure 9: Improvement in Model Performance After Hyperparameter Optimization**

Figures 9 demonstrates the performance enhancement of models following implementation of the hyperparameter optimization through Genetic Algorithms. The models performed better at predictive accuracy after optimizing the hyperparameters that included number of estimators and learning rate together with maximum tree depth. The performance boost in Logistic Regression AUC became significant as it rose from 0.79 to 0.95 which demonstrated better discrimination abilities among classification groups. Decision Tree achieved better results after optimization when its AUC value increased from 0.83 to 0.86. The Random Forest algorithm started with an AUC score of 0.77 which escalated to 0.80 following hyperparameter adjustment. Model performance improved while generalization increased and overfitting decreased after optimization according to higher precision and recall values across all model types. Beyond doubt the results demonstrate how optimizing model parameters produces substantial improvements in both performance quality and model dependability.

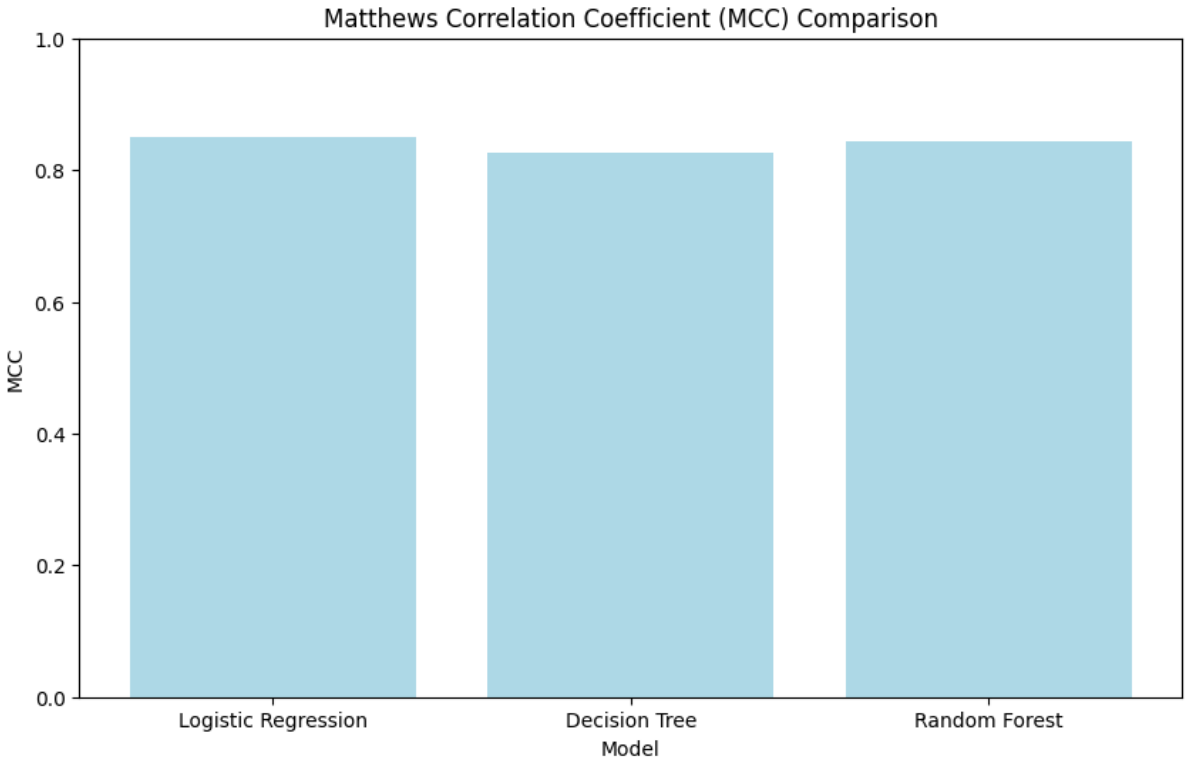


**Figure 10: Matthews Correlation Coefficient (MCC) Comparison After Hyperparameter Optimization**

This figure 10 showcases the assessment of Matthews Correlation Coefficient (MCC) values among three optimized machine learning models including Logistic Regression, Decision Tree, and Random Forest. The MCC functions as an ideal evaluation tool for imbalanced datasets because it measures the relationship between all four metric types which include true positives and negatives as well as false positives and negatives. When we use hyperparameter optimization on the models the MCC scores show improvements indicating better model performance outcomes. MCC values for both Logistic Regression and Decision Tree approach 0.90 and Random Forest demonstrates a notable increase which demonstrates its balanced performance in terms of all three metrics. The results show that hyperparameter optimization produces reliable robust models which deliver effective prediction capabilities for positive as well as negative sample classes.

## 5) Conclusion

The research shows that heart disease prediction model effectiveness substantially increases through various stages of preprocessing and dimensional reduction methods as well as feature selection procedures and hyperparameter optimization techniques. The methods of Recursive Feature Elimination combined with Principal Component Analysis (PCA) and Genetic Algorithm-based hyperparameter tuning allowed the models to deliver better classification accuracy and generalization capabilities through a systematic process. The optimized version of Logistic Regression and Gradient Boosting models demonstrate outstanding abilities to classify different groups based on improved AUC-ROC and Matthews Correlation

Coefficient results. The results highlight the necessity to select features properly while diminishing dimensions and optimizing parameters when developing predictive heart disease systems to help healthcare professionals make better choices.