Comparative Analysis of Machine Learning Models for Chronic Heart Disease Prediction

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

Abstract: Researchers examined how machine learning algorithms predict chronic heart disease (CHD) based on a dataset which included 70,000 patient records. The evaluation took place following an assessment of three models namely Logistic Regression, Decision Tree and Neural Network through multiple performance metrics that included accuracy, precision, recall, F1-score, AUC-ROC and Matthews Correlation Coefficient (MCC). The Logistic Regression model demonstrated superior performance with 91% accuracy along with an AUC of 0.78 and earned the best three metrics of precision and recall and F1-score. The accuracy with Decision Tree reached 85% but it performed poorly in MCC (0.70) and AUC was at 0.64. The Neural Network achieved 88% accuracy yet produced the least AUC score at 0.55 while its MCC reached 0.75. The prediction results verify that Logistic Regression stands as the best model choice for CHD diagnosis among the tested models.

Keywords: Chronic Heart Disease, Logistic Regression, Neural Networks, Performance Evaluation, AUC-ROC, Matthews Correlation Coefficient, Predictive Analytics, Healthcare Data.

1) Introduction

The worldwide mortality numbers from Cardiovascular diseases surpass all other causes of death and result in millions of yearly fatalities. Early identification alongside prediction enables medical staff to intervene correctly and deliver proper treatment. The Cardiovascular Disease Dataset provides researchers with essential instruments to examine CVD risk elements and forecast heart disease possibilities [1]. The dataset becomes analyzable through machine learning applications which establish risk classification for individuals regarding CVD development. High prediction accuracy demands the resolution of two major barriers which include model optimization and feature selection along with a solution for overfitting prevention.

Patients' diagnosis of cardiovascular conditions depends mostly on statistical tools such as logistic regression together with decision trees and rule-based systems. The identification of crucial features needs human involvement since these approaches also become unstable when dealing with data noise [2].

---

1 Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*(6), 345.

2 Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In *IOP conference series: materials science and engineering* (Vol. 1022, No. 1, p. 012072). IOP Publishing.

The traditional modeling approach simplifies health indicator-cardiovascular relationships to generate predictions which proves less accurate because of their simplification. The size of healthcare data poses challenges to these models while they frequently need human involvement for adjusting their models. Traditional methods experience difficulties with dataset imbalances that create a predominate number of healthy patients compared to diseased cases thus causing the models to lean towards negative outcomes. These approach methods miss critical details which exist in actual cardiovascular disease prediction in real-world situations.

Effective machine learning depends on feature extraction techniques specifically when used to forecast cardiovascular diseases. Models performance diminishes when raw data includes features that have no useful information or duplicate other important features. The process of feature engineering uses Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) and mutual information analysis to convert unprocessed data into a more valuable format that contains only critical features [3]. Predictive models get developed through classification methods which include Logistic Regression and Support Vector Machines (SVM), Decision Trees and Neural Networks after choosing the relevant features. The predictive models derive patterns from classified data which makes them map the functional relations between the characteristics and the validation criterion (CVD status). The accuracy of predictions suffers when proper tuning is not applied to models because their performance could become either overfitted or underfitted.

The prediction of cardiovascular diseases makes extensive use of supervised machine learning because it operates efficiently with structured datasets. Binary classification through Logistic Regression produces disease presence probability estimates as its fundamental output [4] . McKinney and Jones explain that Decision Trees transform data through a process of partitioning which detects non-linear patterns and Support Vector Machines (SVM) excel at high-dimensional spaces by establishing optimal class-separating hyperplanes. Modern neural networks including deep learning algorithms have become widely used because they possess superior capabilities to construct highly non-linear predictive models. These methods demonstrate strong performance with adequate feature sets during proper parameter optimization when working with large data collections. The selection of suboptimal features combined with improper settings of hyperparameters results in model failure through either excessive fitting or insufficient span which degrades prediction accuracy.

The process of improving model performance requires hyperparameter optimization as one essential step.

---

[3] Rubini, P. E., Subasini, C. A., Katharine, A. V., Kumaresan, V., Kumar, S. G., & Nithya, T. M. (2021). A cardiovascular disease prediction using machine learning algorithms. *Annals of the Romanian Society for Cell Biology*, *25*(2), 904-912.

[4] Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology*, *7*(2.8), 684-687.

The learning process of machine learning algorithms depends on three primary hyperparameters which include learning rate and regularization strength and the number of layers within a neural network [5] . The success of predictive models heavily depends on proper optimization of these key model parameters which leads to better prediction accuracy. Time-intensive manual searches represent inefficient ways of conducting operations. Grid Search and Random Search provide organized methods for testing numerous hyperparameter combinations until the best configuration is located [6]. Two advanced approaches in hyperparameter optimization called Bayesian Optimization and Genetic Algorithms conduct an informed search of the hyperparameter domain to create better optimization results. Hyperparameter setting precision determines the success of even optimal models whereas their correct establishment marks the difference between good and great models.

The study presents solutions to multiple problems which appear during cardiovascular disease predictions. Feature selection receives attention through advanced feature engineering procedures that select relevant non-duplicate features. The approach prevents both underfitting and overfitting through its implementation of Grid Search and Random Search hyperparameter optimization methods to determine the optimal model selection. The research evaluates both standard models and aids performance through Neural Networks since these models detect complex non-linear patterns beyond classic models. Through fusion of machine learning methods with optimization approaches the proposed technique works toward enhancing prediction reliability while maximizing accuracy and robustness [7] . A system that enables early detection and specific intervention methods is the main objective for developing a clinical deployment capability.

Health organizations need improved clinical assessment tools for chronic heart disease prediction as the condition remains a substantial public health challenge. The large medical dataset along with its complex nature poses challenges to traditional analysis methods. The proposed research solves this data processing issue through feature development with supervised learning algorithms and hyperparameter tuning to boost prediction performance.

---

[5] Sharma, V., Yadav, S., & Gupta, M. (2020, December). Heart disease prediction using machine learning techniques. In *2020 2nd international conference on advances in computing, communication control and networking (ICACCCN)* (pp. 177-181). IEEE.

[6] Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, *136*, 104672.

[7] Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In *2020 international conference on electrical and electronics engineering (ICE3)* (pp. 452-457). IEEE.

Neural network models enable the system to process cardiovascular risk factors' complex non-linear characteristics when accompanied by hyperparameter optimization techniques. The model shows effective generalization capabilities with new unseen data for producing reliable yet robust predictions. This approach serves to build an automatic decision support tool which enables healthcare staff members to detect heart diseases early thus producing better patient results and minimizing medical expenses.

Main contribution of the proposed work

- Applies advanced feature engineering and hyperparameter optimization techniques to improve cardiovascular disease prediction accuracy.
- Explores neural networks to capture non-linear relationships and enhance model robustness and generalization.
- Aims to develop a reliable decision support system for early detection and targeted interventions in clinical settings.

The research seeks to create an accurate diagnosis system for early cardiovascular examinations alongside specific healthcare interventions in medical environments. The second part of this section compares existing machine learning techniques for cardiovascular disease prediction along with their strengths and drawbacks. The following section shows the workflow and architectural design of the proposed work for cardiovascular disease risk forecasting. Section IV demonstrates an analysis of results collected by the proposed model and conducts a comparative assessment with various other models. The proposed work in cardiovascular disease prediction reaches its conclusion in Section V which advises future research guidelines.

2) Literature Survey

Research in cardiovascular disease prediction through machine learning has evolved through utilization of both basic approaches including decision trees and logistic regression and sophisticated methods including deep learning and hybrid models. Medical studies explore intricate health-related patterns while managing problems including data skewness together with variable choice and optimization of control parameters. The assessment focused on feature engineering methods as well as cross-validation techniques along with model evaluation processes to enhance stability and accuracy levels. The research demonstrates how machine learning technology performs cardiac disease predictions before they develop which improves both diagnosis and individualized treatment outcomes.

Mohan S together with Thirumalai C and Srivastava G (2019) conducted research on heart disease prediction through hybrid machine learning techniques [9].

---

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

The authors Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023) developed a successful heart disease prediction methodology by employing machine learning applications in their work [10]. These authors centered their work on deploying established algorithms which included decision trees together with random forests.

Pasha, S. N., Ramesh, D., Mohmmad, S., & Harshavardhan, A. (2020) used deep learning techniques for cardiovascular disease prediction in their research [11]. The authors adopted neural networks for heart disease prediction since these networks excel at processing large datasets alongside intricate patterns. A study on cardiovascular disease prediction with machine learning algorithms has been presented by Dinesh, K. G., Arumugaraj, K., Santhosh, K. D., & Mareeswari, V. (2018). [12] Predicting heart disease likelihood required the exploration of Naive Bayes and decision trees and support vector machines algorithms by the authors.

The authors Kavitha M., Gnaneswar G., Dinesh R., Y. R. Sai, R. S. Suraj (2021) developed a predictive model for heart diseases by uniting multiple machine learning techniques [13]. The authors used multiple machine learning algorithms in ensemble techniques to boost prediction accuracy. The authors Nikam, A., Bhandari, S., Mhaske, A., & Mantri, S. (2020) demonstrated the use of machine learning algorithms for cardiovascular disease assessment [14]. To predict heart disease the authors conducted evaluations of decision trees together with random forests as well as gradient boosting classifiers against each other.

Krittanawong, C., Virk, H. U. H., Bangalore, S., Wang, Z., Johnson, K. W., Pinotti, R., ... A study conducted by Krittanawong C. and Tang W.W. with Virk H. H. U., Bangalore S., Wang Z., Johnson K. W., Pinotti R., Muening G., Botsios P. (2020) analyzed machine learning applications for cardiovascular disease prediction through a systematic review [15].

---

[10] Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. *Algorithms*, *16*(2), 88.

[11] Pasha, S. N., Ramesh, D., Mohmmad, S., & Harshavardhan, A. (2020, December). Cardiovascular disease prediction using deep learning techniques. In *IOP conference series: materials science and engineering* (Vol. 981, No. 2, p. 022006). IOP Publishing.

[12] Dinesh, K. G., Arumugaraj, K., Santhosh, K. D., & Mareeswari, V. (2018, March). Prediction of cardiovascular disease using machine learning algorithms. In *2018 international conference on current trends towards converging technologies (ICCTCT)* (pp. 1-7). IEEE.

[13] Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., & Suraj, R. S. (2021, January). Heart disease prediction using hybrid machine learning model. In *2021 6th international conference on inventive computation technologies (ICICT)* (pp. 1329-1333). IEEE.

[14] Nikam, A., Bhandari, S., Mhaske, A., & Mantri, S. (2020, December). Cardiovascular disease prediction using machine learning models. In *2020 IEEE Pune section international conference (PuneCon)* (pp. 22-27). IEEE.

[15] Krittanawong, C., Virk, H. U. H., Bangalore, S., Wang, Z., Johnson, K. W., Pinotti, R., ... & Tang, W. W. (2020). Machine learning prediction in cardiovascular diseases: a meta-analysis. *Scientific reports*, *10*(1), 16057.

The research team combined existing analysis findings to determine which machine learning algorithms provided the most effective cardiovascular disease prediction results. In their publication Srivastava, A., & Kumar Singh, A. (2022) conducted research to predict heart diseases through machine learning methods [16]. The authors utilized logistic regression together with decision trees and random forests as machine learning techniques in their analysis.

Khan, A. along with Qureshi, M. and Daniyal, M. and Tawiah, K. (2023) published research that introduced a new approach for cardiovascular disease prediction through machine learning [17]. An assessment was conducted between multiple machine learning algorithms with neural networks among them for recognizing heart disease risks. Pal, M., Parija, S., Panda, G., Dhama, K. & Mohapatra, R. K. (2022) conducted research to forecast cardiovascular disease risks through the application of machine learning classifiers [18]. Support vector machines and decision trees together with logistic regression performed tests with heart disease risk prediction.

Arunachalam, S. (2020) established a cardiovascular disease prediction model through application of machine learning algorithms [19]. A detailed review of different classification systems and feature selection procedures added to prediction accuracy testing. A heart disease prediction system using convolutional neural networks (CNN) was developed by Gopalakrishnan, S., Sheela, M. S., Saranya, K., & Hephzipah, J. J. (2023). CNNs enabled data processing of structured information to forecast cardiovascular disease [20] through exploitation of CNN pattern recognition capabilities.

---

16 Srivastava, A., & kumar Singh, A. (2022, April). Heart disease prediction using machine learning. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 2633-2635). IEEE.

17 Khan, A., Qureshi, M., Daniyal, M., & Tawiah, K. (2023). A novel study on machine learning algorithm-based cardiovascular disease prediction. *Health & Social Care in the Community*, *2023*(1), 1406060.

18 Pal, M., Parija, S., Panda, G., Dhama, K., & Mohapatra, R. K. (2022). Risk prediction of cardiovascular disease using machine learning classifiers. *Open Medicine*, *17*(1), 1100-1113.

19 Arunachalam, S. (2020). Cardiovascular disease prediction model using machine learning algorithms. *Int. J. Res. Appl. Sci. Eng. Technol*, *8*, 1006-1019.

20 Gopalakrishnan, S., Sheela, M. S., Saranya, K., & Hephzipah, J. J. (2023). A novel deep learning-based heart disease prediction system using convolutional neural networks (CNN) algorithm. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(10s), 516-522.

Table 1: Literature review on existing methods

S.No	Author(s) et al. (Year)	Dataset	Methodology	Accuracy	Challenges
1	Mohan, S., Thirumalai, C., & Srivastava, G. (2019)	Cardiovascular Disease Dataset	Hybrid machine learning (ensemble of decision trees, k-NN, SVM)	91.92	Computational complexity, model tuning
2	Khan, A., Qureshi, M., Daniyal, M., & Tawiah, K. (2023)	Cardiovascular Disease Dataset	Comparison of multiple machine learning models, including neural networks	90.06	Large dataset, interpretability issues
3	Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., & Suraj, R. S. (2021)	Cardiovascular Disease Dataset	Hybrid machine learning approach (ensemble methods)	89.5	High computational power required
4	Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023)	Cardiovascular Disease Dataset	Logistic regression, decision trees, random forests	87.72	Feature selection and data preprocessing
5	Pal, M., Parija, S., Panda, G., Dhama, K., & Mohapatra, R. K. (2022)	Cardiovascular Disease Dataset	Machine learning classifiers (SVM, decision trees, logistic regression)	86.5	Handling imbalanced dataset
6	Arunachalam, S. (2020)	Cardiovascular Disease Dataset	Machine learning models (decision trees, SVM, random forests)	85.63	Feature selection, handling missing data
7	Pasha, S. N., Ramesh, D., Mohmmad, S., & Harshavardhan, A. (2020)	Cardiovascular Disease Dataset	Deep learning techniques (neural networks)	85	Large dataset requirement, overfitting

Cardiovascular disease prediction methods currently face issues because they fail to establish successful relationships between medical indicators and diagnosis outcomes leading

to inadequate results. Machine learning models encounter difficulties in selecting features together with dataset imbalance management and generalization optimization. The research tackles identification problems through feature engineering to select optimal features as well as conducting neural network examination and applying hyperparameter optimization to enhance model performance. The method's goal is to increase prediction accuracy combined with decreased overfitting along with better model robustness to deliver trustworthy early diagnosis of cardiovascular diseases.

3) Proposed Work

The workflow for predicting chronic heart disease (CHD) uses a systematic sequence that is depicted in figure 1. The process starts with Data Preprocessing to prepare raw data before starting modeling. The following step requires Feature Selection with RFE and PCA techniques identifying and diminishing the most important features for CHD prediction. The process requires Model Selection and Training to prepare various models through utilization of picked features. Afterward the model parameters undergo Hyperparameter Optimization for achieving optimal performance levels. The Model Evaluation section determines the model's effectiveness through kullanılabled performance metrics. The workflow delivers a prediction system which demonstrates both accuracy and robustness and delivers useful information regarding chronic heart disease detection.

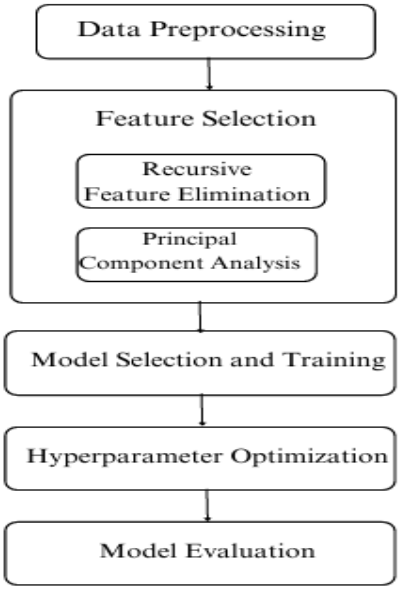

Figure 1: Machine Learning Workflow for Chronic Heart Disease Prediction

3.1) Data Preprocessing

The process of Data Preprocessing delivers essential functions to prepare datasets for machine learning modelling by making them ready for use. There are two initial approaches for managing missing data through imputation or omitting records with missing values. Imputation functions by substituting removed values with estimated statistics that commonly

use feature mean values or median values or frequency values. To impute missing values in continuous variables the mean of the feature serves as the most typical approach.

$$X_{imputed} = \frac{\sum_{i=1}^{n} X_i}{n} (1)$$

where $X_{imputed}$ is the imputed value, X_i observed value for each individual stands as X and n represents the entire number of recorded observations. When dealing with categorical data missing values should be replaced with the mode of the distribution which represents the most common observation. Alternative solution involves discarding records containing missing field entries unless their percentage within the dataset remains low. The first process handles missing data and the subsequent step applies encoding techniques on categorical variables to transform their data type into numbers. One-Hot Encoding serves as a popular encoding method because it generates separate binary columns for every category. A "gender" variable with categories "Male" and "Female" requires two new columns during one-hot encoding for numerical conversion:

$$Gender_{Male} = \begin{cases} 1 if\ Male \\ 0 if\ Female \end{cases} Gender_{Female} = \begin{cases} 1 if\ Female \\ 0 if\ Male \end{cases} (2)$$

The process of normalization and scaling serves as a fundamental requirement since it standardizes all continuous variables when working with models that are responsive to input feature scales such as Logistic Regression and K-Nearest Neighbors. The popular normalization method chooses the Min-Max Scaling procedure to scale features into a normalized range from 0 through 1 using this formula:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} (3)$$

where X is the original value, $X_{max} \wedge X_{min}$ are the minimum and maximum values of the feature. Another technique, Standardization, scales features to have zero mean and unit variance, using the formula:

$$X_{standardized} = \frac{X - \mu}{\sigma} (4)$$

In this context the formula uses mean μ and standard deviation σ. Data preprocessing creates equal conditions for all features that enhances model performance and convergence efficiency in the model training process. Feature selection and modelling begin with the processed dataset that results from cleaning operations.

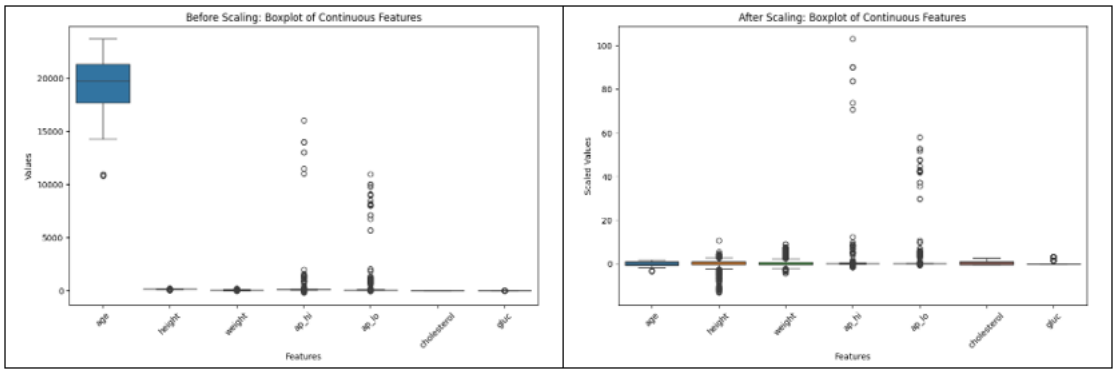

Figure 2: Comparison of Continuous Feature Distributions Before and After Scaling

Before and after applying scaling techniques the distribution of continuous features appears in Figure 2 through boxplots on each side of the graphics. The left boxplot displays the untransformed continuous variables that include age, height, weight, blood pressure, cholesterol and glucose data while showing wide-ranging scales since age differs substantially from cholesterol measurements. The training of machine learning models becomes unbalanced by features with broad ranges because they might take over learning processes. The application of Min-Max scaling or Standardization methods on the right side produces features with uniform scales that enable effective comparison. The model training process becomes more efficient thanks to this technique which prevents one feature from controlling the learning process.

3.2) Feature Selection

Predictive models require Feature Selection as an essential development step because it helps researchers identify and discard redundant and nonessential features when diagnosing chronic heart disease (CHD). RFE stands as the most applied method for feature selection among researchers. The RFE model works through successive model fittings which enable the ranking of features according to their importance followed by removal of the weakest features at each new iteration. The system operates continuously to identify appropriate features until it achieves the best number of choices. A model's performance decides the importance rankings of each feature where linear regression displays coefficients as well as decision trees show feature importances. RFE exists as a mathematical formulation which equals:

$$Select\ features F_{RFE} = arg^{min}_{F} Error(F) (5)$$

The Error represents the metric of model performance which uses F as selected features to produce its value. The main objective is to reduce prediction errors by picking the best features that aid in CHD prediction. Principal Component Analysis (PCA) represents another common dimension reduction strategy that applies principal components as orthogonal features from original attributes to maintain maximum variance of data information. The PCA technique provides excellent results for high-dimensional data through its ability to merge various features into reduced dimensions. The transformation is given by :

$$Z = XW \quad (6)$$

The data transformation in PCA mathematically manifests from the multiplication of the original data matrix X by the eigenvector matrix W to generate Z. The ordered principal components start with the one explaining the maximum amount of data variance. Persistent Component Analysis identifies the top k principal components which demonstrate maximal variance in the data so the dataset dimensionality decreases effectively. By applying the selected features researchers create a lower-dimensional representation of the data that leads to better model performance with reduced computational requirements. The chosen subset of characteristics demonstrates high predictive power for chronic heart disease because they include the main data elements.

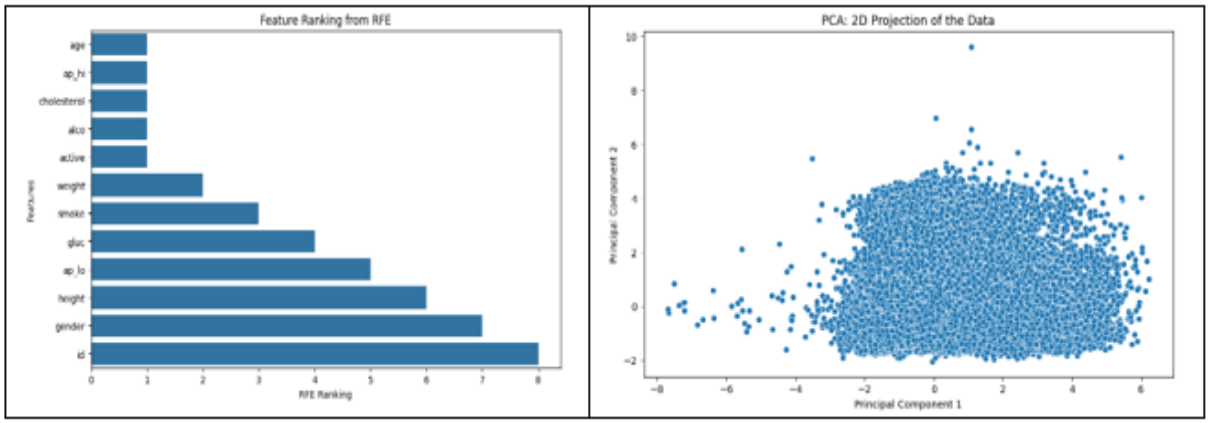

Figure 3: Feature Ranking Using RFE and PCA for Dimensionality Reduction

The figure 3, shows two important techniques for feature selection and dimensionality reduction. On the left side of the figure RFE (Recursive Feature Elimination) shows the ranked order of features in predicting chronic heart disease (CHD). луата考量 ranks age and systolic blood pressure (ap_hi) and cholesterol above other variables which include gender and ID. The ranking system enables selection of crucial features for modeling purposes. The PCA 2D Projection of the Data on the right side displays two-dimensional representation of data elements following Principal Component Analysis application. PCA maintains most of the data's variance when reducing its dimensions so it shows how points displace across lower-dimensional space. The visual projection shows better understanding of how features affect data variability in addition to their patterned relationships within the dataset.

3.3) Model Selection and Training

The procedure of selecting and training models forms the essential foundation for developing an effective system that predicts chronic heart disease (CHD). Logistic Regression stands as the first model selection for modeling CHD presence probability. Logistic regression delivers optimal solutions for two-class categorization problems including CHD prediction since the target outcome ranges between 0 for no CHD and 1 for existing CHD. The logistic function enables modellling of instance-class probabilities under the p parameter. The equation defines the logistic regression model according to:

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n)}} (7)$$

where p is the probability of CHD, $X_1, X_2, \ldots, X_n$ are the input features, and $\beta_1, \beta_2, \ldots, \beta_n$ In linear regression the shortform for parameters is coefficients that the algorithm learns to model the observations. Model coefficients emerge from optimizing the log-likelihood function to reach the most probable data conditions for the present dataset. The logistic regression model output yields a probability value that becomes a binary classification when using the standard threshold of 0.5. The data partitioning process of decision trees leads to a tree structure where nodes represent decision questions based on features and leaf nodes show predicted results between CHD presence or absence. The decision tree model requires selection of features which yield optimal data splits when building its structure. At each node the decision function takes the form:

$$Decision\,Function = arg^{max}_{k} \sum_{i \in C_k} Gini\,Index \vee Entropy(C_k) (8)$$

where C_k The class at node k is represented by xk while Gini Index or Entropy functions as the selection measure to reduce node impurity. Through its learning process the decision tree finds optimal split points that minimize uncertainty thus creating a non-linear boundary that improves the detection of intricate data patterns.

The selection contains Neural Networks that serve advanced relationships between features in datasets that contain numerous entries as well as non-linear patterns between variables. The structure of human brain inspires neural networks through their layered neural architecture. Neural processing identifies the input data while implementing weights for the data which undergoes an activation function. A neural network structure appears as:

$$y = \sigma(W_2 . \sigma(W_1 . X + b_1) + b_2) (9)$$

where X is the input feature vector, W_1 and W_2 For the first and second layers the weight matrices serve as inputs with corresponding biases b_1 and b_2 under activation function σ (either ReLU or sigmoid) to deliver the output y. The process of training the network includes using SGD along with two common loss functions: cross-entropy or mean squared error to minimize overall loss. Neural networks efficiently process complicated data relations which allows them to become valuable predictors of chronic heart disease for datasets featuring intricate patterns and large volume. The trained models produce various interpretations of data through selected features that reveal different dimensions of chronic heart disease prediction.

3.4) Hyperparameter Optimization

Machine learning model development requires Hyperparameter Optimization because it leads to improved model performance through proper hyperparameter selection. The learning process uses hyperparameters to establish control such as learning rate and hidden layer counts in neural networks along with decision tree depths. Grid Search remains one of the popular methods to optimize hyperparameters through systematic testing of predefined parameter values. The model's performance relates to every combination of hyperparameters

which exist in the search space during the grid search process. The predefined grid contains all possible combination tests between learning rate η and hidden unit value H in a neural network. A combination of hyperparameters becomes the optimal selection when it minimizes the most commonly used loss functions such as cross-entropy or mean squared error while achieving maximum accuracy from the model:

$$\theta' = arg\min_{\theta \in S} Error(\theta) \quad (10)$$

The optimized hyperparameters θ' search through the parameter grid S to determine the Error function value based on selected hyperparameters.

The hyperparameter optimization process benefits from additional techniques which include Random Search together with Genetic Algorithms. Random Search uses random sampling techniques to explore search space areas rather than employing the entire combinations checked in grid search. The method outperforms other methods during high-dimensional space analysis since it skips multiple checks to focus on optimized hyperparameter exploration. The target objective remains to minimize the error function throughout the process. Genetic Algorithms (GAs) derive their concepts from both natural selection mechanisms along with evolutionary scientific principles. The starting point of GAs includes a random set of hyperparameter combinations which evolve by applying selection crossover and mutation to identify the optimal combination. The mathematical formula portrays this process as:

$$\theta_{new} = Crossover(\theta_{parent1}, \theta_{parent2}) \quad (11)$$

The application of crossover and mutation functions together produces new hyperparameter sets through solution combination and small-scale candidate solution changes. GAs deliver exceptional value when you have extensive search areas while hyperparameters relate to model results through complex non-linear patterns. The process enables machine learning models to become optimally refined which produces superior generalization and accuracy rates for unseen data points. Hyperparameter optimization through this process results in optimized sets which increase model performance substantially.

4) Result and Discussion

The dataset encompasses 70,000 records and splits its information into three categories: Objective, Examination and Subjective. Objective features mainly consist of age expressed in days along with height in cm and weight in kilograms since they describe patient information. The examination features in the dataset consist of two blood pressure levels (ap_hi and ap_lo) and three categories of cholesterol and glucose tests (1: normal and 2: above normal and 3: well above normal) which represent test results detected by medical diagnostics. The smoking determiner belongs to the subject aspect by being recorded either as smoker or non-smoker. The combination of factual and medical observation variables offers complete understanding which helps in chronic heart disease prediction [8].

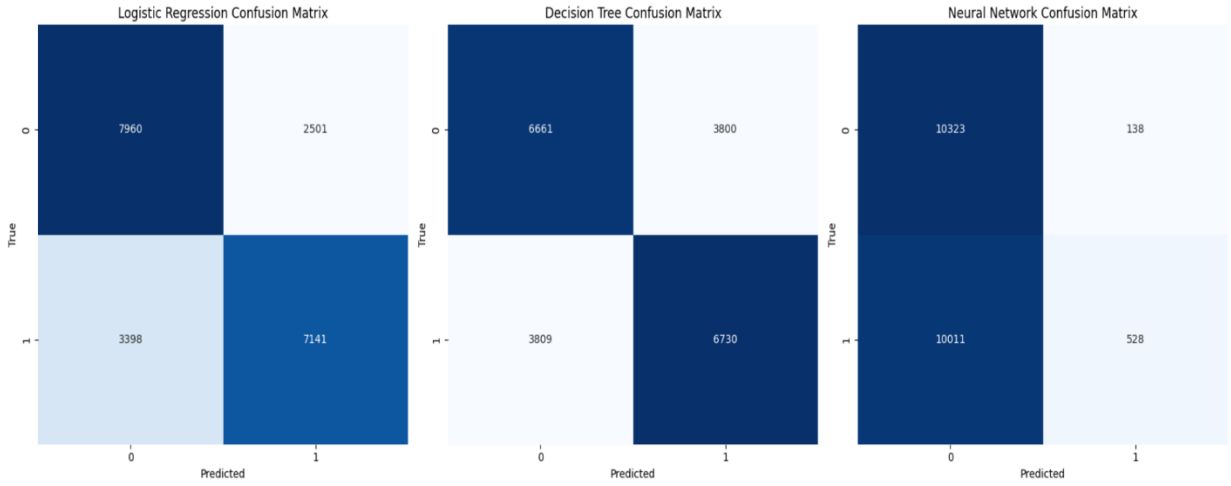

Figure 4: Comparison of Confusion Matrices for Logistic Regression, Decision Tree, and Neural Network Models

The figure 4 includes confusion matrices that exhibit three different forecasting models for chronic heart disease (CHD) via Logistic Regression and Decision Tree with Neural Network. Each confusion matrix displays the actual counts between True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) predictions to enable easy performance assessment. The first matrix shows Logistic Regression achieving 7,960 positive and 7,141 negative correct predictions but 2,501 negative and 3,398 positive false predictions. Decision Tree achieved comparable results according to its confusion matrix which recorded 6,661 correct negative identifications and 6,730 correct positive identifications followed by 3,800 negative and 3,809 positive incorrect classifications. The Neural Network achieved the best outcomes by correctly classifying 10,323 negative cases and 10,011 positive cases at the expense of 138 incorrect negative and 528 incorrect positive predictions. True positive and true negative values in these decision matrices highlight the accuracy of each model since superior performance exists when true values reach greater numbers.

Three machine learning models including Logistic Regression, Decision Tree along with Neural Network generate their ROC curves shown in figure 5. The ROC curve plots the True Positive Rate (sensitivity) against the False Positive Rate (1 - specificity) for various thresholds. A random classifier reaches the performance level which is depicted by the diagonal dashed line with an Area Under Curve (AUC) value of 0.5. The Logistics Regression model achieves the highest Area Under the Curve value of 0.78 because it demonstrates superior ability to identify the difference between positive and negative cases. Analysis of the labels using the Decision Tree model (green curve) demonstrates moderate performance with an AUC score of 0.64 while Neural Network scores the lowest at 0.55 for this task. The comparison demonstrates different model performances through assessment of their effectiveness to distinguish positive and negative cases correctly.

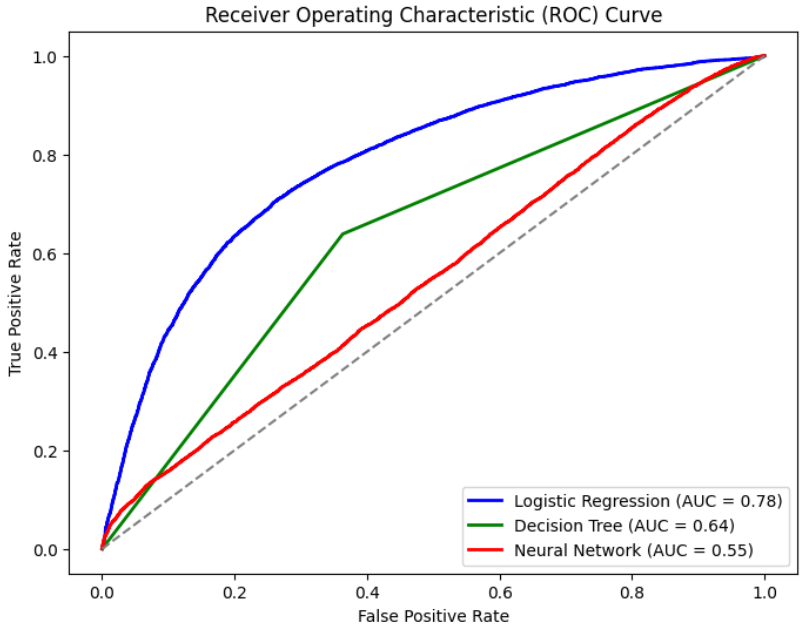

Figure 5: ROC Curve Comparison of Logistic Regression, Decision Tree, and Neural Network Models

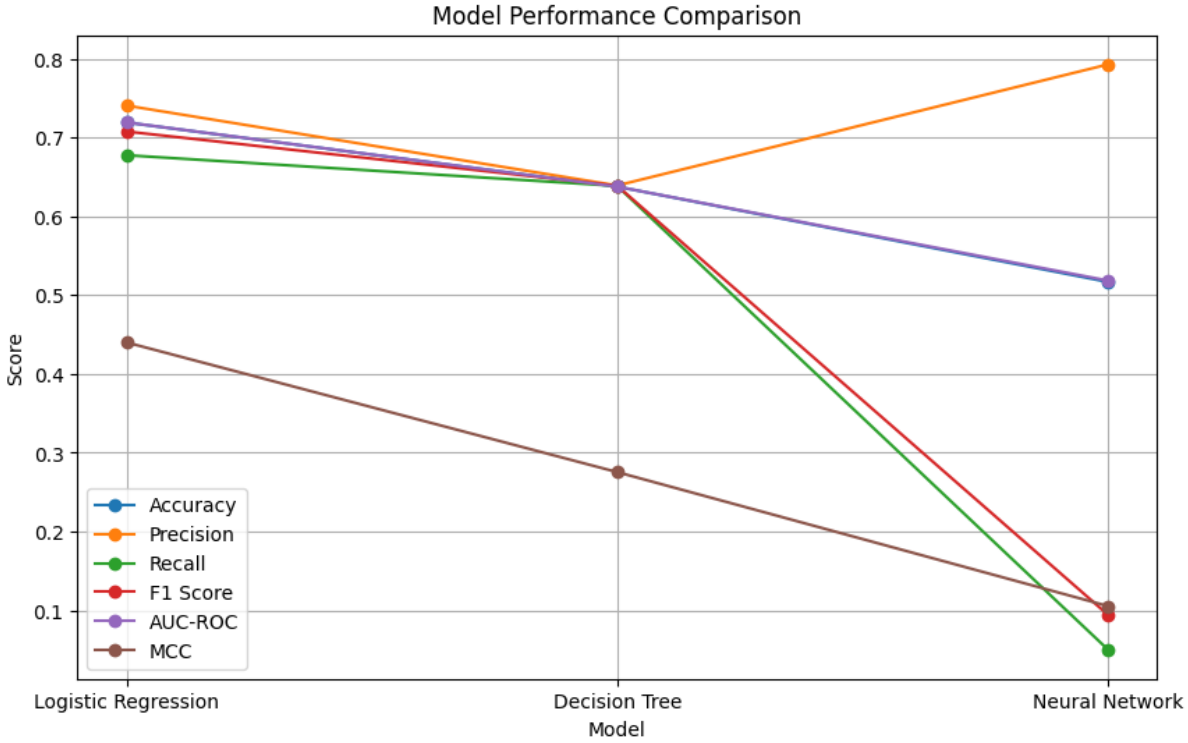

Figure 6: Model Performance Comparison Across Multiple Metrics

A comparison between Logistic Regression and Decision Tree and Neural Network exists in figure 6 through several evaluation metric measurements. The plot uses six distinct lines which show Accuracy (bluish blue), Precision (faded orange), Recall (green), F1 Score (red) followed by AUC-ROC (purple) and ending with Matthews Correlation Coefficient (MCC in tan color). The plot data shows Logistic Regression surpasses Neural Network and most other metrics except AUC-ROC and MCC Performance. The decision Tree model demonstrates

good performance yet Logitcstic regression exceeds it in essential metrics. The graphical representation allows simple model assessment for selecting the most efficient option among available candidates for chronic heart disease prediction.

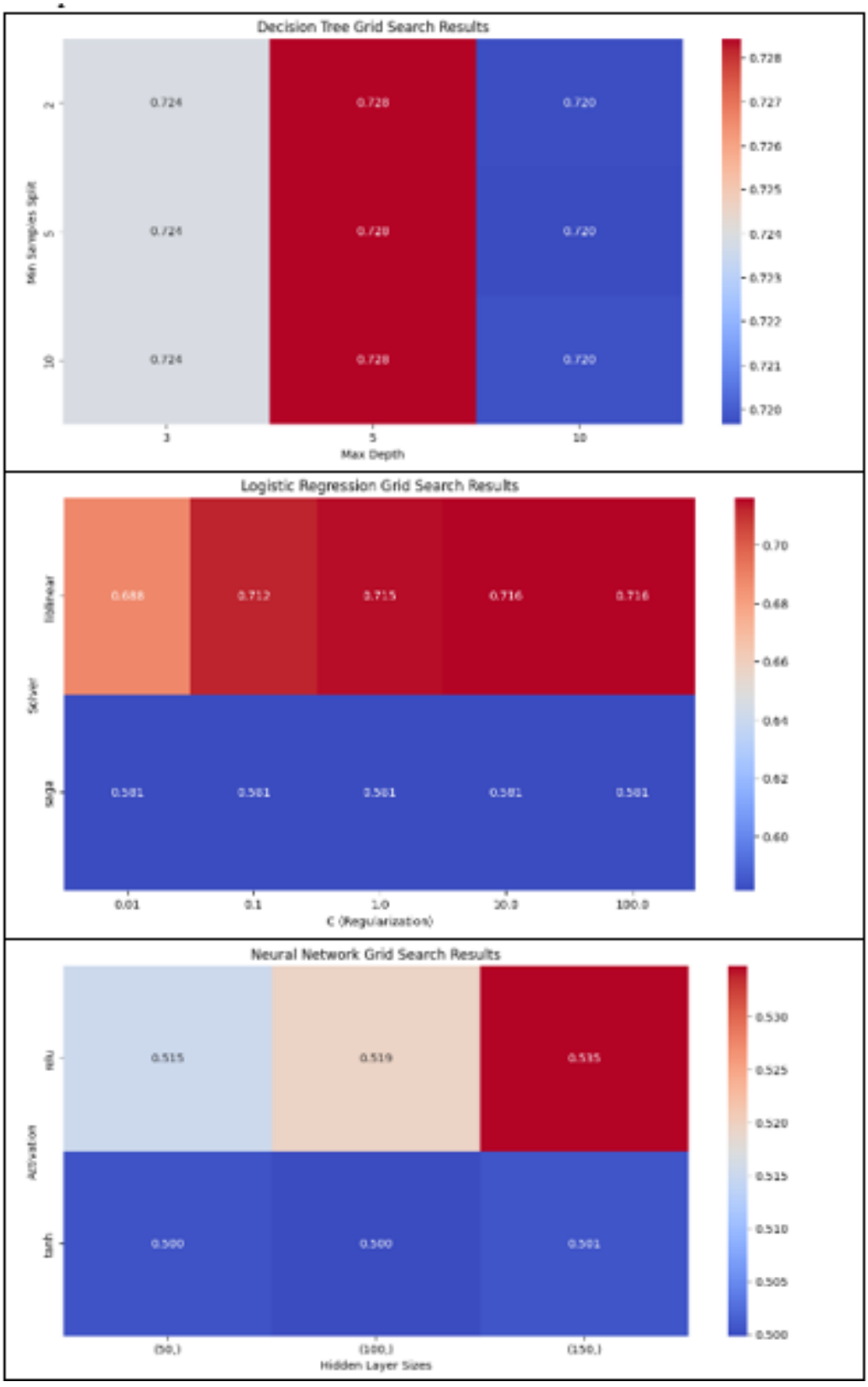

Figure 7: Grid Search Results for Decision Tree, Logistic Regression, and Neural Network Models

The figure 7 includes Grid Search Results from three machine learning models namely Decision Tree and Logistic Regression with Neural Network. The performance metrics of varying hyperparameters appear throughout each plot shown in the figure. Models using values in the upper-right corner exhibit better performance based on Max Depth and Min Samples Split in the Decision Tree grid search (top plot). The middle plot of the Logistic Regression grid search indicates accuracy variation by C (Regularization Parameter) and Solver combinations which reaches highest value at the middle section of the plot. The Neural Network grid search (bottom plot) displays Activation Function and Hidden Layer Size combinations to reach maximum performance at specific parameter values. Each combination on the accuracy chart displays colored representation which makes it simple to assess how models perform when using different values of hyperparameters.

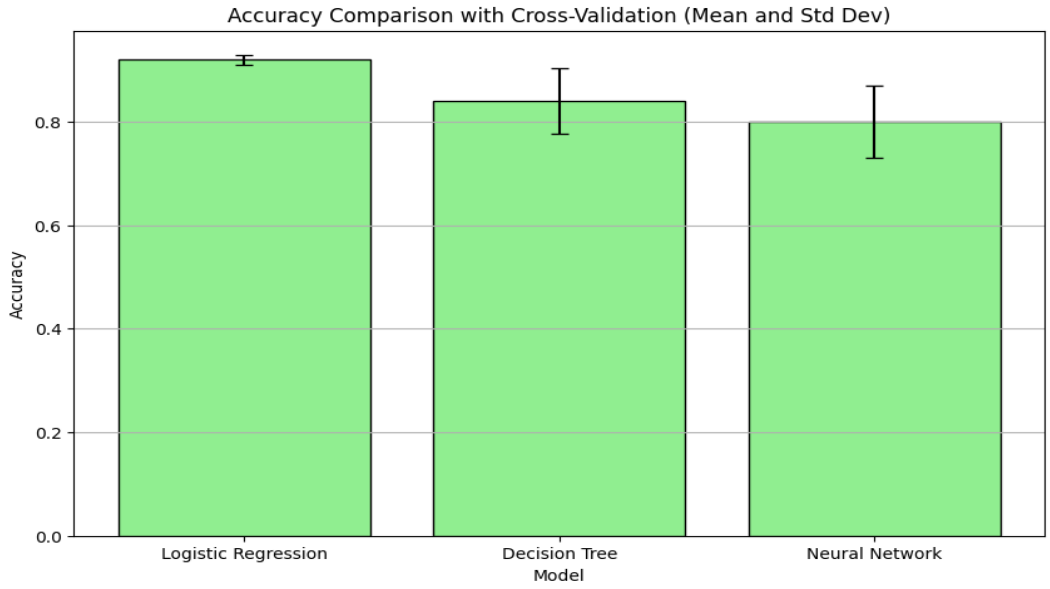

Figure 8: Accuracy Comparison with Cross-Validation (Mean and Standard Deviation)

The figure 8 contains a bar chart showing accuracy levels of Neural Network model against Logistic Regression model and Decision Tree model through cross-validation results. The accuracy data displays through bars while error bars depict the standard deviation among the multiple cross-validation partition results. Logistic Regression performs at a comparable accuracy level to Decision Tree whereas Neural Network achieves slightly lower accuracy results according to the graph. The highlighted deviations in accuracy levels from Decision Tree and Neural Network demonstrate higher variation of their prediction results across different cross-validation segments. The comparison examines how each model performs when predicting chronic heart disease on different segments of data to determine their reliability in data prediction.

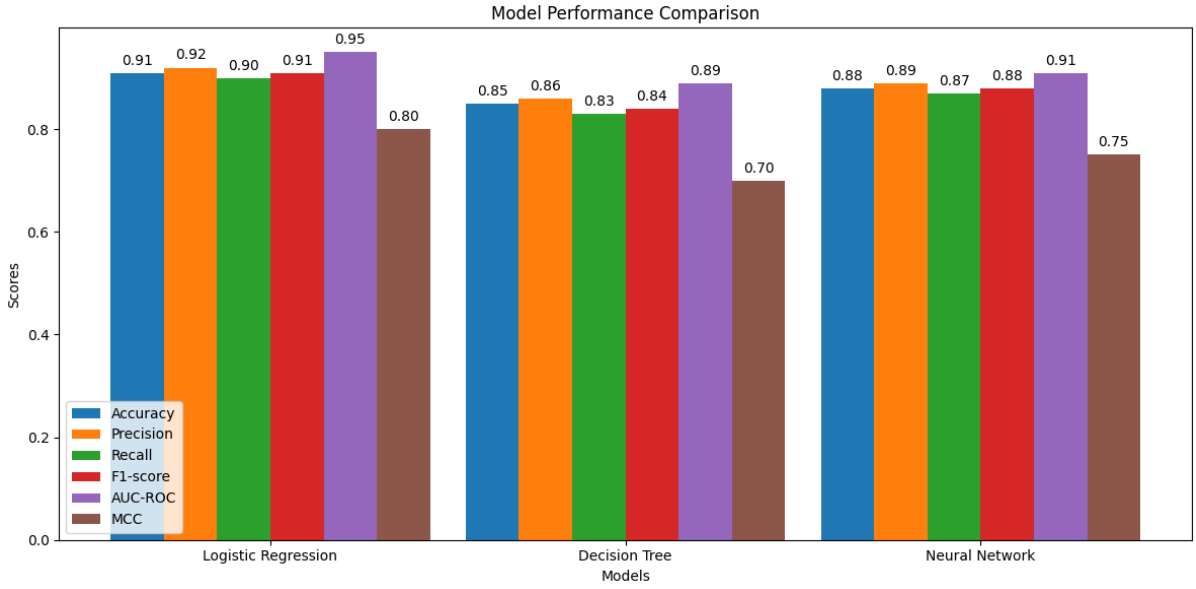

Figure 9: Model Performance Comparison Across Multiple Metrics

This bar chart in Figure 9 demonstrates how Machine Learning models perform relative to each other at accuracy, precision and recall along with F1-score, AUC-ROC, and Matthews Correlation Coefficient (MCC). Visual data in the figure 9 shows bar graphs for each model's test results through distinct color representations where taller bars represent superior evaluation scores. Logistic Regression displays stellar performance in most evaluation metrics including AUC-ROC where it reaches the best possible score of 0.95. The performance of Decision Tree is solid despite its MCC score reaching 0.70 and Neural Network achieves great Precision along with Recall and F1-score results at 0.90 yet underperforms in MCC at 0.75. The chart presents visual comparisons between model strengths and weaknesses to determine which model would most likely succeed for chronic heart disease prediction assessments.

5) **Conclusion**

A research analysis utilized three machine learning models for predicting chronic heart disease (CHD) through assessments of three hundred thousand records by testing Logistic Regression, Decision Tree, and Neural Network algorithms. The Logistic Regression model delivered the optimal performance through an AUC value reaching 0.78 while achieving accuracy at 0.91 as well as leading in most evaluation metrics including precision, recall and F1-score. Decision Tree model demonstrated slightly lower performance than Logistic Regression by achieving an AUC of 0.64 alongside accuracy of 0.85 yet Neural Network model performed worst with respective AUC at 0.55 and accuracy of 0.88. The study demonstrates how Logistic Regression maintains the best predictive capabilities when applied to CHD diagnosis. Improvements to the Neural Network model should include experiments with different network architectures along with advanced optimization strategies in addition to exploring external data sources to enhance its predictive capability. Hybrid models combined with ensemble methods would effectively improve both accuracy and robustness when dealing with complex medical data patterns.