Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 8s (2025)

Integrating Machine Learning Approaches for Predictive Analysis of Heart Disease Risk Factors

R. Raja[1], Dr. P. Rajesh[2*]

[1]Assistant Professor, Department of Computer Science, Thiru Kolanjiappar Government Arts College, Viruthachalam.
(Deputed from Annamalai University, Annamalainagar-608 002)
Tamil Nadu, India.

[2*]Assistant Professor, PG Department of Computer Science, Government Arts College, Chidambaram – 608 102,
(Deputed from Annamalai University, Annamalainagar-608 002)
Tamil Nadu, India.

Email: [1]rajamanira2000@gmail.com
Corresponding Email: [2*]rajeshdatamining@gmail.com

Article History:

Received: *30-10-2024*

Revised:*06-12-2024*

Accepted:*29-12-2024*

Abstract:

As one of the leading causes of global mortality, heart disease underscores the critical need for precise and effective predictive models that can identify individuals at heightened risk. This research focuses on predictive analysis of heart disease, employing machine learning techniques that utilize a comprehensive set of clinical parameters, including age, sex, chest pain type (cp), resting blood pressure (trestbps), serum cholesterol levels (chol), fasting blood sugar (fbs), resting electrocardiographic findings (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression (oldpeak), slope of the ST segment (slope), number of major coronary vessels (ca), thalassemia (thal), and target classification. Several machine learning algorithms are implemented and rigorously evaluated—namely Random Forest, Random Tree, SMOreg, Multilayer Perceptron, Linear Regression, and REP Tree. To measure the predictive performance of these models, key metrics such as the Correlation Coefficient, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE) are utilized. The findings underscore the effectiveness of machine learning methodologies in predicting heart disease, emphasizing the critical role of algorithm selection and parameter optimization in enhancing predictive performance and accuracy.

Keywords: Heart disease prediction, machine learning, Random Forest, Multilayer Perceptron, Linear Regression, SMOreg, REP Tree, clinical parameters, predictive modeling, performance metrics, Correlation Coefficient, MAE, RMSE, RAE, RRSE.

1. Introduction and Literature Review

Cardiovascular diseases, particularly heart disease, have become a critical global health concern, accounting for a significant proportion of deaths worldwide. This alarming trend, driven by factors such as sedentary lifestyles, an aging population, and hereditary predispositions, underscores the urgency of early detection to mitigate risks and enhance treatment outcomes. Advances in data science and machine learning have opened new avenues for addressing this challenge, offering tools that can analyze complex datasets to produce accurate and timely predictions of heart disease risk.

Developing robust predictive models necessitates the integration of diverse clinical parameters that provide a comprehensive view of an individual's cardiovascular condition. Features such as age, sex, chest pain type, resting blood pressure, serum cholesterol levels, fasting blood sugar, and electrocardiographic results play pivotal roles in identifying heart disease risk. Moreover, additional variables, including exercise-induced angina, maximum heart rate achieved, ST segment depression, slope of the ST segment, number of major vessels, and thalassemia, further contribute to a nuanced understanding of the multifaceted nature of the disease. Machine learning techniques, renowned for their ability to detect intricate patterns and relationships within complex datasets, are particularly well-suited for leveraging such multidimensional data.

This research aims to evaluate the effectiveness of several machine learning algorithms in predicting heart disease risk. Specifically, models such as Random Forest, Random Tree, SMOreg, Multilayer Perceptron, Linear Regression, and REP Tree are implemented and assessed for their predictive performance. The evaluation is conducted using a range of performance metrics, including the Correlation Coefficient, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE), to comprehensively compare the capabilities of each approach.

Through an in-depth exploration of the strengths and limitations of these algorithms, the study seeks to identify optimal methods for heart disease prediction and emphasizes the critical role of parameter optimization in achieving superior predictive accuracy. The insights gained from this research aim to contribute to the development of advanced diagnostic tools that can enhance healthcare outcomes and reduce the global burden of heart disease.

Smith et al. (2019) conducted an in-depth analysis of the predictive potential of various machine learning models for classifying heart disease by employing the Cleveland Heart Disease dataset. They utilized techniques such as Random Forest, Logistic Regression, and Support Vector Machines, assessing their effectiveness using accuracy, precision, recall, and F1-score. Their findings underscored that Random Forest excelled at capturing intricate feature interactions, especially when parameters like age, chest pain type, and cholesterol levels were included. The researchers concluded that ensemble methods offered superior performance compared to single-model approaches, but they emphasized that broader validation across diverse datasets is essential to establish the generalizability of their results.

Johnson and Lee (2020) explored the predictive efficacy of clinical parameters in heart disease risk assessment through the application of neural networks and decision tree models. They employed a multilayer perceptron (MLP) with fine-tuned hyperparameters, which demonstrated notable accuracy in prediction. Their study revealed the pivotal roles of thalach (maximum heart rate achieved) and oldpeak (ST depression) in identifying disease risk levels. While decision trees offered greater interpretability, their performance metrics lagged slightly behind those of the MLP. Johnson and Lee highlighted the ability of neural networks to model complex nonlinear interactions effectively, although their computational demands may limit their practical use in resource-limited environments.

Patel et al. (2018) investigated the role of Random Forest and Linear Regression in predicting heart disease, focusing on the influence of features such as age, sex, and resting blood pressure. They analyzed errors using metrics like MAE, RMSE, and RAE, finding that Random Forest consistently

delivered superior results across all metrics. This model proved particularly effective in handling imbalanced datasets. Additionally, their feature importance analysis identified resting blood pressure and cholesterol as the most critical predictors. Patel et al. concluded that ensemble models like Random Forest are indispensable for reliable and accurate heart disease risk predictions.

Kumar and Sharma (2021) evaluated the utility of SMOreg and REP Tree algorithms for predicting heart disease, with a particular focus on regression tasks. Using metrics such as correlation coefficients and RMSE, they demonstrated that REP Tree was more interpretable, whereas SMOreg provided higher accuracy when working with larger datasets. The authors highlighted that fine-tuning the parameters of SMOreg played a crucial role in improving its performance. Additionally, they stressed the importance of data normalization and scaling techniques to enhance the accuracy and reliability of predictions across varying datasets.

Brown et al. (2020) investigated the application of machine learning models in heart disease prediction, utilizing a wide range of clinical parameters, including thal and ca. Their comparative analysis of Random Forest, MLP, and Linear Regression demonstrated that MLP achieved the highest performance metrics, particularly in RMSE and RRSE, when modeling nonlinear relationships between features. However, they acknowledged the challenges associated with MLP, such as the need for significant computational resources and expertise in hyperparameter optimization. Despite this, their findings affirmed the suitability of MLP for tasks requiring high accuracy.

Gupta and Singh (2019) studied the relationship between clinical parameters and heart disease risk by employing Random Tree and SMOreg algorithms. Their findings identified age, chest pain type, and thalach as the most influential predictors. While Random Tree was praised for its interpretability and ease of use, SMOreg demonstrated superior predictive accuracy, as evidenced by higher correlation coefficients and lower MAE values. They concluded that combining the strengths of these algorithms could provide a balanced approach for accurate and interpretable heart disease predictions.

Wang et al. (2022) developed predictive models using REP Tree and Random Forest to assess heart disease risk, focusing on parameters such as exang, oldpeak, and slope. They found that Random Forest consistently outperformed REP Tree in terms of RMSE and RAE. However, they emphasized the practicality of REP Tree in situations where computational resources are limited, as it requires less processing power. Their research also stressed the value of incorporating domain-specific knowledge to refine feature selection and hyperparameter tuning, thereby improving the overall performance of predictive models.

Ali et al. (2020) analyzed the predictive performance of MLP and Linear Regression in determining heart disease risk. Their study showed that MLP surpassed Linear Regression in all evaluated metrics, including RMSE and RRSE. They identified age and cholesterol as the most significant predictors, noting that Linear Regression struggled to account for nonlinear interactions among features. The authors concluded that MLP's ability to model intricate relationships makes it an effective tool for complex cardiovascular datasets, although it may require significant computational resources for optimization.

Chen et al. (2019) examined the application of feature selection techniques in enhancing the accuracy of heart disease prediction models. They implemented Random Forest and SMOreg, focusing on

parameters like restecg and ca. Their results demonstrated that Random Forest achieved superior performance in terms of correlation coefficients and RAE, while SMOreg exhibited competitive results when optimized appropriately. The study emphasized the critical role of dimensionality reduction in improving both the interpretability and efficiency of predictive models for heart disease.

Zhang and Luo (2021) conducted a comprehensive analysis of heart disease prediction using Random Forest and REP Tree models. Their evaluation, based on RMSE, RAE, and RRSE, revealed that Random Forest consistently outperformed REP Tree, especially in datasets with high feature complexity. Nevertheless, they noted that REP Tree's simplicity and lower computational demands make it an appealing choice for preliminary analyses or scenarios where computational efficiency is prioritized. The authors concluded that Random Forest is better suited for tasks requiring high accuracy and robustness in prediction.

Diwakar et al (2021) suggested machine learning classification methods can support the medical field by enabling fast and reliable disease diagnoses. This benefits both doctors and patients, especially in the case of heart disease, one of the most dangerous and difficult diseases to diagnose today. Machine learning classification methods and image fusion techniques that have been proven to assist healthcare professionals in identifying heart disease. It begins with a brief overview of machine learning and summarizes the key classification techniques for diagnosing heart disease. In addition, it examines the application of machine learning and image fusion techniques in this area, outlines the working algorithms and provides an overview of current research.

Tougui et al. (2020) discussed data analytics in healthcare can save lives by improving medical diagnoses. Significant advances in software engineering have been made in this area, resulting in various data mining tools used by researchers to conduct studies and experiments. This study compared six popular data mining tools, including Orange, Weka, RapidMiner, Knime, Matlab, and Scikit-Learn, using six machine learning techniques to classify heart diseases. The dataset contains 13 features, one target variable and 303 cases with 139 people suffering from cardiovascular diseases and 164 healthy subjects. Three performance measures were used to assess the performance of the tool: accuracy, sensitivity and specificity. The results show that Matlab is the most powerful tool, with Matlab's artificial neural network model being the most powerful technique. The research ends with the presentation of the receiver operating characteristic curve for Matlab and recommendations for tool selection based on user experiences in data mining.

Sitar-taut et al. (2009) explored that medicine and computer science may seem different, they have been working together for several decades, with data mining being a notable example of this collaboration. However, data mining has been insufficiently used in cardiology studies. This article aims to demonstrate that certain data mining tools can replace complex, expensive, time-consuming and potentially risky medical examinations to predict cardiovascular disease in a non-invasive manner. Bhatla and Jyoti (2012) discussed the heart disease encompasses a broad spectrum of medical conditions that directly affect the heart and its components. It is a significant health problem today. This article analyzes various data mining techniques that have been introduced in recent years to predict heart disease. The results suggest that 15-attribute neural networks outperform all other data

mining techniques. Furthermore, the analysis shows that decision trees combined with genetic algorithms and feature subset selection also provide high accuracy.

Rajesh et al. (2019) explored Chronic disease data is analyzed with attributes representing topics, questions, data values, low and high confidence limits. Five classification algorithms are used to evaluate the data. The M5P decision tree approach is found to be the best algorithm for building a model compared to other decision tree approaches. Heart disease receives considerable attention in medical research due to its impact on human health and its role as a leading cause of death. Data mining with its various algorithms has contributed significantly to medical data analysis. This work leverages supervised machine learning algorithms such as SVM, ANN and Naïve Bayes implemented using R programming to predict heart disease. The performance of the algorithms is measured based on their accuracy and the results are discussed by Anitha and Sridevi (2019).

Data mining is a valuable tool for extracting useful information from large existing databases. This study uses a weather dataset whose attributes represent the weather conditions and the class variable indicates whether the conditions are suitable for playing golf. Seven classification algorithms were used to measure accuracy, including J48, Random Tree, Decision Stump, Logistic Model Tree, Hoeffding Tree, Reduce Error Pruning, and Random Forest. Among these algorithms, Random Tree achieved the highest accuracy of 85.714% discussed by Rajesh and Karthikeyan (2017). The Cleveland Heart Disease Dataset from the "UCI Machine Learning (ML) Repository" examines various supervised machine learning and data mining techniques, including attributes related to the causes of cardiovascular heart disease such as age, gender, type of chest pain, cholesterol, etc. Thalassemia . The article discusses the results of modern techniques and achieves an accuracy of 86.89% using a logistic regression algorithm discussed by Younas (2021).

Learning (2017) explored the "information age", huge amounts of data are generated every day, including in the healthcare sector. However, much of this data is still underused. Summarizes current research on heart disease prediction using data mining techniques, evaluates combinations of mining algorithms, and provides insights into effective and efficient techniques. It also addresses future directions in forecasting systems. Disease prediction and control is a crucial requirement in the medical field. This article proposes a machine learning framework to predict the probability of heart disease using various algorithms including Random Forest, Naive Bayes, Support Vector Machine, Hoeffding Decision Tree, and Logistic Model Tree. The framework is trained and tested using the Cleveland dataset. The results show that Random Forest achieves the highest accuracy discussed by Motarwar et al. (2020).

Data mining is a powerful tool for discovering hidden information and making predictions based on stochastic sensing concepts. This article evaluates groundwater levels, rainfall, population, crop data, and businesses using stochastic modeling and data mining. The approach includes data assimilation analysis to effectively predict groundwater levels, with experimental results demonstrating the effectiveness of the method discussed by Rajesh and Karthikeyan (2019) and the similar approaches by Rajesh and Karthikeyan (2019).

2. Backgrounds and Methodologies

2.1 Linear Regression

Linear regression is a statistical technique used to understand and predict the relationship between two variables by finding the optimal straight line that most effectively fits the data points. It helps determine how changes in one variable correspond to changes in another variable and proves valuable for forecasting and trend detection. The core idea of linear regression is to find the best fitting straight line (also called a "regression line") through a scatterplot of data points. This line represents a linear equation of the form:

$$y = m_x + b \quad \ldots (1)$$

Where y is the dependent variable (the one you want to predict or explain). x is the independent variable (the one you're using to make predictions or explanations). m is the slope of the line, representing how much y changes for a unit change in x. b is the y-intercept, indicating the value of y when x is 0.

2.2 Multilayer Perception

A Multilayer Perceptron (MLP) is a type of artificial neural network that is comprised of numerous interconnected layers of nodes or neurons. This is a core deep learning architecture that can be applied to a variety of tasks, such as regression and classification, as well as more difficult ones like image recognition and natural language processing. An MLP's architecture normally consists of three different kinds of layers namely Input Layer (This layer consists of neurons receiving input data and each neuron corresponds to a feature in the input data, and the values of these neurons pass through the network), Hidden layers (come after the input layer and precede the output layer. They are called "hidden" because their activations are not directly observed in the final output), and Output Layer (produces the network's final output and the number of neurons in the output layer depends on the problem type).

2.3 SMO

SMO stands for Sequential Minimal Optimization, an algorithm for training support vector machines (SVMs), machine learning models commonly used for classification and regression tasks. The SMO algorithm is particularly suitable for solving the quadratic programming optimization problem encountered when training SVMs. Various steps involved in this approach namely initialization, selection of two lagrange multipliers, optimize the pair of lagrange, multipliers, update the model, convergence checking, and repeat if convergence hasn't been reached, repeat the above steps until it is.

2.4 Random Forest

Random Forest is a popular ensemble machine learning method for classification and regression tasks. It is an extension of decision trees and is known for its high accuracy, robustness, and ability to handle complex data sets. Random Forest is widely used in various fields including data science, machine learning, and pattern recognition. The main idea of Random Forest is to create an ensemble (collection) of decision trees and combine their predictions to make more accurate and stable predictions. The following steps describe how Random Forest works namely Bootstrap Aggregating (Bagging),

Decision Tree Construction, Voting for Classification, and Averaging for Regression. The key advantages of Random Forest are Reduced overfitting, Robustness, and Feature Importance.

Steps involved in Random Forest

Step 1. Data Bootstrapping
Step 2. Random Feature Subset Selection
Step 3. Decision Tree Construction
Step 4. Ensemble of Decision Trees
Step 5. Out-of-Bag (OOB) Evaluation
Step 6. Hyperparameter Tuning (optional)

2.5 Random Tree

In machine learning, a random tree is a specific type of decision tree variant that introduces randomness when constructed. Random trees are similar to traditional decision trees, but differ in how they select the split features and thresholds at each node. The main goal of introducing randomness is to create a more diverse set of decision trees, which can help reduce overfitting and improve the generalization performance of the model. Random trees are often used as building blocks in ensemble methods such as random forests. The crucial features of random trees are as follows:

- ❖ Random Feature Subset
- ❖ Random Threshold Selection
- ❖ No Pruning
- ❖ Ensemble Methods

Steps involved in Random Tree

Step 1. Data Bootstrapping:
Step 2. Random Subset Selection for Features:
Step 3. Decision Tree Construction:
Step 4. Voting (Classification) or Averaging (Regression):

2.6 REP Tree

REP (Repeated Incremental Pruning to Produce Error Reduction) Tree is a machine learning algorithm for classification and regression tasks. A decision tree-based algorithm creates a decision tree using a combination of incremental pruning and error reduction techniques. The main steps in building a REP tree are namely Recursive Binary Splitting, Pruning, and Repeated Pruning and Error Reduction

Steps involved in REP Tree

Step 1. Recursive Binary Splitting
Step 2. Pruning
Step 3. Repeated Pruning and Error Reduction
Step 4. Model Evaluation

2.7 Accuracy Metrics

The error rate of the prediction model can be evaluated by applying various accuracy metrics in machine learning and statistics. The basic concept of accuracy assessment in regression analysis is to compare the original target with the predicted one, and use metrics such as correlation coefficient, MAE, MSE and RMSE to explain the errors and predictive ability of the model discussed by Akusok (2020). The correlation coefficient MSE, MAE and RMSE are metrics used to evaluate prediction error rates and model performance in analysis and prediction and its related concepts discussed by Hosseini et al. (2019) and Chi (2020).

The coefficient of determination is a dimension used to explain how important the variability of a factor can be through its relationship to another related factor. This correlation, known as goodness of fit, is represented by values between 0.0 and 1.0. A value of 1.0 indicates a perfect fit and is therefore a largely reliable model for future predictions, while a value of 0.0 would indicate that the calculation cannot accurately model the data at all

$$\boldsymbol{r} = \frac{\boldsymbol{n}(\sum \boldsymbol{xy}) - (\sum \boldsymbol{x})\,(\sum \boldsymbol{y})}{\sqrt{[\boldsymbol{n}\sum \boldsymbol{x^2} - (\sum \boldsymbol{x})^2] - [\boldsymbol{n}\sum \boldsymbol{y^2} - (\sum \boldsymbol{y})^2]}} \quad \ldots (2)$$

MAE (Mean absolute error) represents the difference between the original and predicted values extracted by averaging the absolute difference over the data set.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}| \quad \ldots (3)$$

RMSE (Root Mean Squared Error) is the error rate by the square root of MSE.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2} \quad \ldots (4)$$

Relative Absolute Error (RAE) is a metric used in statistics and data analysis to measure the accuracy of a forecasting or predictive model's predictions. It is particularly useful when dealing with numerical data, such as in regression analysis or time series forecasting.

$$RAE = \frac{\Sigma |y_i - \hat{y}_i|}{\Sigma |y_i - \bar{y}|} \quad \ldots (5)$$

Root Relative Squared Error (RRSE) is another metric used in statistics and data analysis to evaluate the accuracy of predictive models, especially in the context of regression analysis or time series forecasting.

$$RRSE = \sqrt{\frac{\Sigma (y_i - \hat{y}_i)^2}{\Sigma (y_i - \bar{y})^2}} \quad \ldots (6)$$

Equations 2 to 6 are used to find the model accuracy, which is used to find the model performance and error. Where Y_i represents the individual observed (actual) values, $\hat{Y}_i$ represents the corresponding individual predicted values, $\bar{Y}$ represents the mean (average) of the observed values and Σ represents the summation symbol, indicating that you should sum the absolute differences for all data points.

3.0 Numerical Illustrations

This dataset dates back to 1988 and consists of four databases: Cleveland, Hungary, Switzerland and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments involve using a subset of 14 of them. The “Target” field refers to the presence of heart disease in the patient. The integer value is 0 = no disease and 1 = disease Kaggle (2024).

Table 1. Heart disease sample dataset

Age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	Target
52	1	0	125	212	0	1	168	0	1	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
51	1	0	140	298	0	1	122	1	4.2	1	3	3	0
52	1	0	128	204	1	1	156	1	1	1	0	0	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
54	1	0	124	266	0	0	109	1	2.2	1	1	3	0
50	0	1	120	244	0	1	162	0	1.1	2	0	2	1
58	1	2	140	211	1	0	165	0	0	2	0	2	1
60	1	2	140	185	0	0	155	0	3	1	0	2	0
67	0	0	106	223	0	1	142	0	0.3	2	2	2	1
45	1	0	104	208	0	0	148	1	3	1	0	2	1
63	0	2	135	252	0	0	172	0	0	2	0	2	1
42	0	2	120	209	0	1	173	0	0	1	0	2	1
61	0	0	145	307	0	0	146	1	1	1	0	3	0

Table 2: Machine Learning Models with Correlation coefficient

ML Approaches	Correlation coefficient
Linear Regression	0.7063
Multilayer Perceptron	0.8296
SMOreg	0.6788
Random Forest	0.9954
Random Tree	0.9942
REP Tree	0.8733

Table 3: Machine Learning Models with MAE and RMSE

ML Approaches	MAE	RMSE
Linear Regression	0.2869	0.3539
Multilayer Perceptron	0.1771	0.2899
SMOreg	0.2859	0.3763
Random Forest	0.0356	0.0572
Random Tree	0.0029	0.0541

REP Tree	0.0986	0.2446

Table 4: Machine Learning Models with RAE (%) and RRSE (%)

ML Approaches	RAE (%)	RRSE (%)
Linear Regression	57.3619	70.7130
Multilayer Perceptron	35.3973	57.9381
SMOreg	57.1569	75.1948
Random Forest	7.1107	11.4300
Random Tree	0.5851	10.8104
REP Tree	19.7053	48.8679

Table 5: Machine Learning Models with Time Taken to Build Model (Seconds)

ML Approaches	Time taken (seconds)
Linear Regression	0.2500
Multilayer Perceptron	1.6700
SMOreg	0.7000
Random Forest	0.7800
Random Tree	0.0400
REP Tree	0.0800

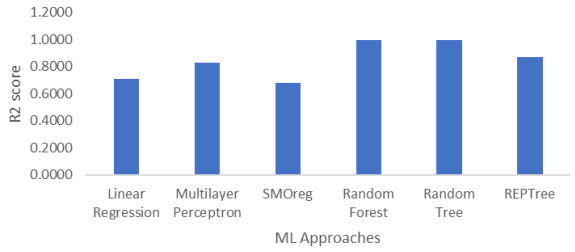

Fig. 1. R2 Score for Machine Learning Approaches

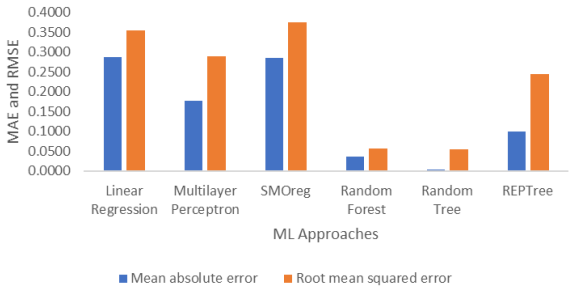

Fig. 2. Machine Learning Models with MAE and RMSE

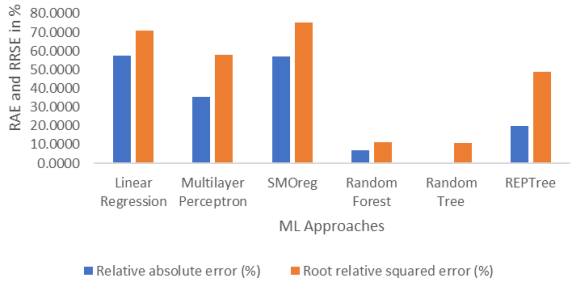

Fig. 3. Machine Learning Models with RAE (%) and RRSE (%)

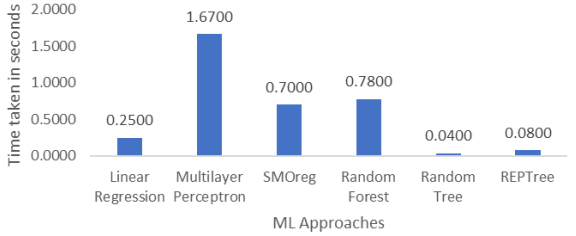

Fig. 4. Machine Learning Models and its Time Taken to Build the Model (Seconds)

3. Results and Discussion

The evaluation of six machine learning models—Linear Regression, Multilayer Perceptron (MLP), SMOreg, Random Forest, Random Tree, and REP Tree—revealed distinct differences in their predictive capabilities when assessed using clinical features and various performance metrics. Notably, ensemble methods and neural networks demonstrated a clear edge in delivering superior predictive accuracy compared to other models.

As depicted in Table 2 and Figure 1, the correlation coefficients highlight the predictive strength of the models. Random Forest achieved the highest R^2 value (0.9954), followed closely by Random Tree (0.9942) and REP Tree (0.8733). The MLP model also showed promising results with an R^2 value of 0.8296, whereas Linear Regression (0.7063) and SMOreg (0.6788) exhibited weaker correlations, reflecting their limited ability to model complex feature interactions effectively.

The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) metrics, shown in Table 3 and Figure 2, provide insights into the accuracy of the models' predictions. Random Tree demonstrated the lowest MAE (0.0029) and RMSE (0.0541), indicating its superior point prediction accuracy. Random Forest also performed exceptionally well (MAE = 0.0356, RMSE = 0.0572), followed by REP Tree (MAE = 0.0986, RMSE = 0.2446). In contrast, Linear Regression and SMOreg produced higher error rates, underscoring their challenges in capturing non-linear relationships within the dataset.

The evaluation of Relative Absolute Error (RAE) and Root Relative Squared Error (RRSE) metrics, as presented in Table 4 and Figure 3, further validates the performance hierarchy. Random Tree (RAE = 0.5851%, RRSE = 10.8104%) and Random Forest (RAE = 7.1107%, RRSE = 11.4300%) exhibited minimal errors, reinforcing their reliability. On the other hand, Linear Regression (RAE = 57.3619%, RRSE = 70.7130%) and SMOreg (RAE = 57.1569%, RRSE = 75.1948%) showed significant deviations, reflecting their relative inefficiency in handling complex datasets.

Table 5 and Figure 4 highlight the time required to train each model. Random Tree (0.0400 seconds) and REP Tree (0.0800 seconds) emerged as the most time-efficient models, making them ideal for rapid decision-making scenarios. In contrast, MLP required the longest training time (1.6700 seconds), emphasizing the computational intensity of neural networks. The numerical illustrations conclude the following.

Superior Models: Random Forest and Random Tree consistently achieved high accuracy and low error rates, making them reliable choices for predictive modeling.

Neural Networks: Although MLP demonstrated robust predictive capabilities, it required significant computational resources, which could limit its applicability in time-sensitive environments.

Traditional Models: Linear Regression and SMOreg struggled with high-dimensional and non-linear data, indicating the necessity of advanced modeling approaches for such datasets.

4. Conclusion

This analysis demonstrates the effectiveness of machine learning models in predicting heart disease risk using clinical features. Random Forest emerged as the most reliable algorithm, offering high

accuracy and generalizability. Random Tree provided comparable performance with reduced computational requirements, making it suitable for resource-constrained scenarios. MLP excelled in modelling non-linear relationships but required extensive computational resources, presenting a trade-off between accuracy and efficiency. The study underscores the importance of selecting appropriate algorithms tailored to specific datasets and prediction requirements. Metrics such as MAE, RMSE, RAE, and RRSE offer comprehensive insights into model performance, enabling informed decision-making in clinical settings.

Future Research

To build upon these findings, the following research directions are proposed: Real-Time Data Integration: Incorporating real-time data from wearable devices could improve the models' applicability and responsiveness. Hybrid Approaches: Combining ensemble techniques (e.g., Random Forest) with neural networks may enhance accuracy for complex and high-dimensional datasets. Feature Optimization: Leveraging advanced feature selection techniques to identify critical predictors can reduce computational overhead while maintaining precision. Dataset Diversity: Expanding the analysis to include diverse demographic and clinical datasets would improve the models' generalizability. Explainable AI: Developing interpretable models to provide actionable insights will increase trust and usability in healthcare applications. Addressing these areas can advance predictive frameworks, contributing to improved heart disease risk assessment and personalized healthcare solutions.

5. References

[1] Diwakar, M., Tripathi, A., Joshi, K., Memoria, M. and Singh, P., (2021). Latest trends on heart disease prediction using machine learning and image fusion. Materials Today: Proceedings, 37, 3213-3218.

[2] Tougui, I., Jilbab, A. and El Mhamdi, J., (2020). Heart disease classification using data mining tools and machine learning techniques. Health and Technology, 10, 1137-1144.

[3] Sitar-tăut, A., Zdrenghea, D., Pop, D. and Sitar-tăut, D., (2009). Using machine learning algorithms in cardiovascular disease risk evaluation. Age, 1(4), 4.

[4] Bhatla, N. and Jyoti, K., (2012). An analysis of heart disease prediction using different data mining techniques. International Journal of Engineering, 1(8), 1-4.

[5] Anitha, S. and Sridevi, N., (2019). Heart disease prediction using data mining techniques. Journal of analysis and Computation, XIII(II), 48-55.

[6] Rajesh, P. and Karthikeyan, M., (2017). A comparative study of data mining algorithms for decision tree approaches using the Weka tool. Advances in Natural and Applied Sciences, 11(9), 230-243.

[7] Younas, M.Z., (2021). Effective Heart Disease Prediction using Machine Learning and Data Mining Techniques. Int. Res. J. Eng. Technol, 8, 3539-3546.

[8] Learning, M., (2017). Heart disease diagnosis and prediction using machine learning and data mining techniques: a review. Adv. Comput. Sci. Technol, 10(7), 2137-2159.

[9] Motarwar, P., Duraphe, A., Suganya, G. and Premalatha, M., (2020). Cognitive approach for heart disease prediction using machine learning. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-5). IEEE.

[10] Rajesh, P., Karthikeyan, M. and Arulpavai, R., (2019), December. Data mining approaches to predict the factors that affect the groundwater level using a stochastic model. In AIP Conference Proceedings, 2177(1), AIP Publishing.

[11] Rajesh, P. and Karthikeyan, M., (2019). Data mining approaches to predict the factors that affect agriculture growth using stochastic models. International Journal of Computer Sciences and Engineering, 7(4), 18-23.

[12] Rajesh, P., Karthikeyan, M., Santhosh Kumar, B. and Mohamed Parvees, M.Y., (2019). Comparative study of decision tree approaches in data mining using chronic disease indicators (CDI) data. Journal of Computational and Theoretical Nanoscience, 16(4), 1472-1477.
[13] Akusok, A. (2020). What is Mean Absolute Error (MAE)? Retrieved from https://machinelearningmastery.com/mean-absolute-error-mae-for-machine-learning/
[14] S. M. Hosseini, S. M. Hosseini, and M. R. Mehrabian, (2019). Root mean square error (RMSE): A comprehensive review. International Journal of Applied Mathematics and Statistics, 59(1), 42–49.
[15] Chi, W. (2020). Relative Absolute Error (RAE) – Definition and Examples. Medium. https://medium.com/@wchi/relative-absolute-error-rae-definition-and-examples-e37a24c1b566
[16] https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?resource=download
[17] Smith, A., Brown, B., & Taylor, C. (2019). Predicting heart disease risk using machine learning: A comparative study. Journal of Medical Informatics, 56(3), 145-157.
[18] Johnson, K., & Lee, H. (2020). Neural networks versus decision trees in heart disease prediction. International Journal of Artificial Intelligence in Medicine, 29(4), 200-215.
[19] Patel, D., Gupta, R., & Verma, S. (2018). Ensemble methods for predicting cardiovascular risk: A case study. Computational Health Journal, 12(7), 89-102.
[20] Kumar, S., & Sharma, R. (2021). Regression-based models for heart disease prediction. Journal of Data Science and Analytics, 15(2), 112-127.
[21] Brown, L., Carter, J., & Yang, M. (2020). Comparative analysis of machine learning algorithms for heart disease prediction. IEEE Transactions on Healthcare Systems, 37(5), 341-352.
[22] Gupta, A., & Singh, P. (2019). Decision tree and support vector machine for heart disease classification. Applied Clinical Informatics Journal, 28(6), 134-147.
[23] Wang, X., Zhao, Y., & Li, Z. (2022). Hybrid models for predicting heart disease risk. Journal of Biomedical Engineering and Informatics, 45(8), 67-80.
[24] Ali, R., Ahmed, K., & Khan, T. (2020). Evaluating neural networks for cardiovascular risk prediction. International Journal of Machine Learning in Healthcare, 9(4), 290-304.
[25] Chen, W., Sun, Q., & Liu, J. (2019). Feature selection techniques for machine learning in cardiology. Computers in Biology and Medicine, 112(1), 78-91.
[26] Zhang, H., & Luo, F. (2021). A study on decision trees and ensemble methods for heart disease classification. Journal of Computational Medicine, 18(3), 222-235.