APPLIED INFORMATION TECHNOLOGY AND COMPUTER SCIENCE
e-ISSN: 2773-5141

AITCS

Vol. 6 No. 1 (2025) 100-119
https://publisher.uthm.edu.my/periodicals/index.php/aitcs

Utilizing Machine Learning Algorithms for Predicting Cardiovascular Disease

Low Yin Qian[1], Noor Zuraidin Mohd Safar[1*]

[1] *Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, 86400, MALAYSIA*

*Corresponding Author: zuraidin@uthm.edu.my
DOI: https://doi.org/10.30880/aitcs.2025.06.01.006

Article Info

Received: 13 June 2024
Accepted: 8 May 2025
Available online: 30 June 2025

Keywords

Cardiovascular disease, Machine Learning, Logistic Regression. Random Forest, Google Colab, Kaggle

Abstract

Cardiovascular disease (CVD) remains a leading cause of global mortality. Early prediction and intervention are vital for improving outcomes and reducing healthcare costs. This study leverages a comprehensive Kaggle dataset, including clinical, demographic, and lifestyle information, to enhance predictive models for CVD. By employing machine learning algorithms like Logistic Regression and Random Forest, the research aims to construct accurate and interpretable models. The study ensures methodological robustness and reproducibility by utilizing tools such as Google Colab and Microsoft Excel for data manipulation and analysis. The performance of these supervised learning algorithms is compared to identify the most accurate model for predicting CVD. Improved prediction accuracy can facilitate early identification and intervention, potentially lowering CVD incidence and improving patient outcomes. The study's findings could significantly impact healthcare by enabling more targeted and effective preventive measures against CVD.

1. Introduction

Cardiovascular disease (CVD) remained a prominent global health concern, accounting for a substantial portion of worldwide mortality. These diseases encompass a range of conditions affecting the heart and blood vessels, including heart attacks, strokes, and heart failure. The early identification of individuals at risk of developing cardiovascular diseases is crucial for implementing timely interventions and preventive measures. In this context, machine learning, a subset of artificial intelligence, offers a promising avenue to harness the power of data for accurate CVD risk prediction. This research project, titled "Machine Learning for Cardiovascular Disease Prediction," aims to build reliable models that can predict an individual's risk for cardiovascular illnesses by utilizing Python-based machine learning techniques. By leveraging various datasets that include clinical measurements and medical histories, the project seeks to develop personalized healthcare solutions and enhance patient outcomes using suitable algorithms.

One of the primary challenges addressed in this research is the complexity and diversity of data related to cardiovascular diseases. This data includes a vast array of clinical variables, genetic factors, and medical histories, making it difficult to effectively analyze and integrate these multifaceted sources to build accurate predictive models. Additionally, interpretability is a significant concern when applying machine learning to healthcare. It is essential to develop models that not only predict cardiovascular disease risk but also provide insights into the underlying factors influencing those predictions. Achieving model and algorithm interpretability within Python-based frameworks is a key challenge of this project. Furthermore, comparing the performance of different

machine learning algorithms across various datasets and parameters is crucial to ensuring high accuracy in predicting the likelihood of cardiovascular diseases.

The objectives of this research include efficiently integrating and preprocessing diverse dataset sources related to cardiovascular diseases, employing machine learning algorithms to develop predictive models for CVD risk, and comparing the performance of different algorithms with various datasets and parameters. The scope of this research involves obtaining keywords related to cardiovascular disease, sourcing datasets from Kaggle, utilizing Google Colab as the primary tool, and applying logistic regression and random forest algorithms.

The expected outcome of this research is the development of accurate machine learning models capable of predicting the risk of cardiovascular diseases in individuals based on diverse data sources, with reasonably high predictive accuracy. This is essential for healthcare professionals to understand and trust the model's decisions. The project holds significant potential to contribute to the early identification of individuals at risk of cardiovascular diseases, enabling timely interventions and preventive measures. This can lead to improved patient outcomes and reduced mortality rates. Moreover, accurately predicting cardiovascular disease risk can support public health initiatives aimed at reducing the prevalence of CVDs, and enhance our understanding of how predictive modeling can be applied to cardiovascular health.

2. Related Work

This section discussed cardiovascular disease, Machine learning and its categories, and the existing research on Machine Learning for cardiovascular disease prediction.

2.1 Cardiovascular Disease

Cardiovascular disease (CVD) includes a range of disorders affecting the heart and blood vessels, posing complex challenges to international health. The global burden of these conditions includes coronary heart disease, cerebrovascular illness, peripheral arterial disease, rheumatic heart disease, congenital heart disease, and thrombotic events such as pulmonary embolism and deep vein thrombosis. Acute events like heart attacks and strokes are primarily caused by blood flow obstructions, which are frequently caused by the build-up of atheromatous plaques in arterial walls [1]. These plaques, which are composed of inflammatory cells and lipids, block arteries that supply the heart or brain, causing ischemia episodes. On the other hand, cerebral hemorrhages or embolic events resulting from clot formation can potentially cause stroke [2]. It is essential to comprehend how clotting processes, arterial plaque development, and acute cardiovascular events interact in order to develop preventative and treatment approaches for these complex and sometimes fatal diseases.

2.2 Machine Learning

A computer science and artificial intelligence (AI) subfield called "machine learning" focuses on using data and algorithms to simulate human learning processes and progressively increase their accuracy. Fundamentally, machine learning is all about developing and using algorithms that help with these choices and forecasts. As they handle more data, these algorithms are built to perform better over time, becoming more precise and efficient [3].

There are 6 processes in the basic machine learning model which are the collection and preparation of data, feature selection, choice of algorithm, selection of models and parameters, training, and performance evaluation [4], as shown in **Fig. 1**.

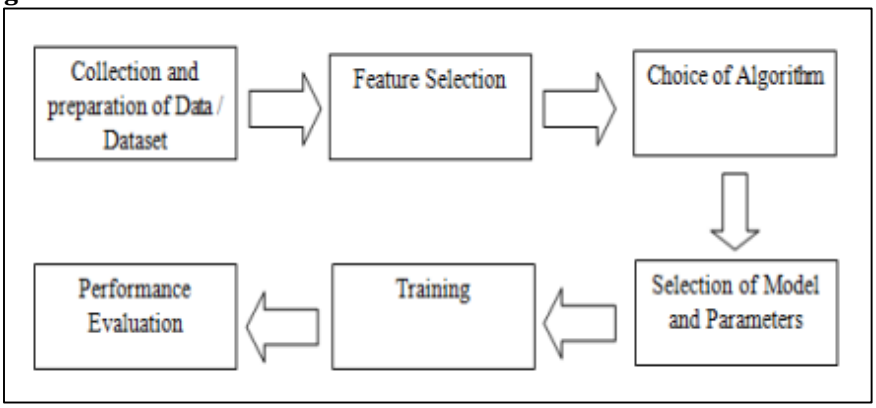


Fig. 1 *Process of Basic Machine Learning Model*

In the process of collection and preparation of data, the data collected contained a lot of both unnecessary and irrelevant data. Thus, the data must be pre-processed in a clean, structured format. As for the feature selection, the irrelevant features in the data obtained will be removed. Next is the choice of algorithm, where the optimal machine learning algorithm is chosen for a specific class issue because not all algorithms are suitable for every

problem. As for the selection of models and parameters, parameters are adjusted to the most suitable values for machine learning algorithms since they require some initial manual intervention. Moreover, during the process of training, only a portion of the dataset is used as training data using the selected model and parameters. Lastly is the process of performance evaluation. The model is evaluated on the learning capacity using various parameters, including accuracy, precision, and recall.

On the other hand, machine learning can be categorized using two types of techniques which are supervised learning and unsupervised learning [3] in **Fig. 2**.

The most prevalent kind of machine learning is supervised learning where a labelled dataset is used to train on the model. During the training phase, the model discovers a mapping between the input (features) and the output (label). The model can predict the result for fresh, untested data once it has been trained. The standard supervised learning algorithms include linear regression for regression problems and logistic regression, decision trees, and support vector machines for classification problems.

On the contrary, unsupervised learning includes using an unlabelled dataset to train the model. It is up to the model to go through the data on its own and identify patterns and relationships. Clustering and dimensionality reduction are frequently applied to this learning process. The model is responsible for grouping similar data points together and minimizing the number of random variables under consideration by applying a collection of primary variables. For example, the k-means algorithm is responsible for clustering problems in unsupervised learning.

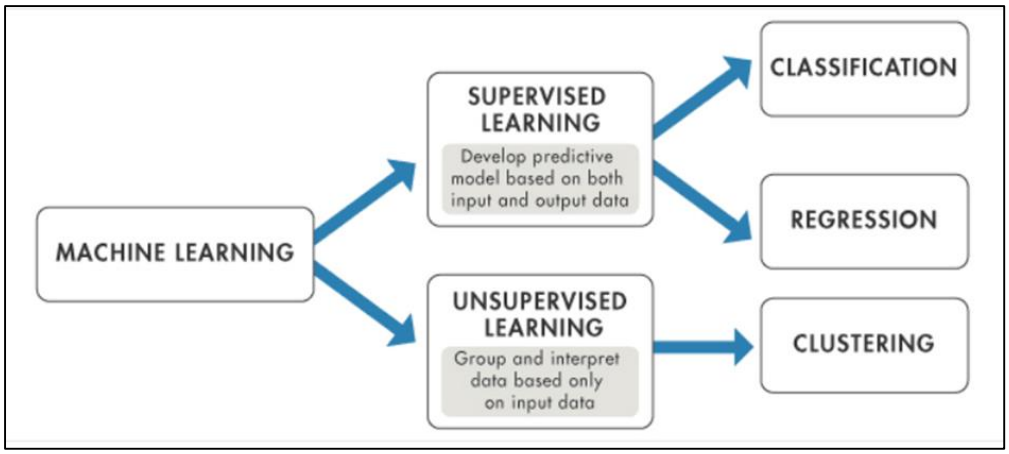


Fig. 2 *Comparison of supervised and unsupervised learning*

2.3 Selected Algorithm

Logistic regression and Random Forest are the supervised algorithms that will be used for this research paper, where both algorithms will be compared to see which is the most accurate in predicting cardiovascular diseases.

2.3.1 Logistics Regression

Logistic Regression is one of the most common approaches for classification in machine learning. Most applications employ the supervised machine learning binary classification technique known as logistic regression. Even though linear regression is similar mainly to logistic regression, the difference between them is that a linear regression graph shows a straight line while a logistic regression graph shows a curve. It works on categorical dependent variables, and the result can be discrete or binary categorical variables 0 or 1 [5].

Logistic regression uses a logistic sigmoid function as a cost function where a predicted real value to a probabilistic value between '0' and '1' is mapped. In Eq. 1, P(x) is the probability estimation function with a value between 0 and 1, x is input to the probability function (algorithm's prediction value), the mathematical constant e is Euler's number and its value is approximately equal to 2.71828 as shown below [5].

$$P(x) = 1 \div \left(1 + e^{-x}\right) \quad \text{(Eq. 1)}$$

Logistic regression offers several advantages, making it a popular choice for binary classification problems. The algorithm is straightforward to implement and requires minimal computational resources. Additionally, it excels when the data is linearly separable and handles scaled input features effectively without the need for extensive tuning. Several assumptions and preprocessing steps must be considered for effective application of logistic regression. The target variable must be binary, with outcomes typically represented by 0 and 1. Independent variables should not exhibit multicollinearity to ensure model accuracy, and the model should include variables that are relevant to the outcome of interest. Furthermore, a sufficient sample size is recommended to achieve reliable results. By adhering to these assumptions and leveraging their advantages, logistic regression remains a robust and interpretable method for binary classification in various applications [6].

Applied Information Technology and Computer Science Vol.6 No.1 (2025) p. 100-119

2.3.2 Random Forest

Random Forest, a machine learning technique employed for both regression and classification tasks, leverages ensemble learning, a methodology that tackles complex problems by amalgamating numerous classifiers. Comprising multiple decision trees, Random Forest undergoes training through bagging or bootstrap aggregating. This process, known as bagging, heightens the accuracy of machine learning models. The algorithm bases its output on predictions from these decision trees. By aggregating the outcomes of multiple trees—typically by averaging or taking the mean—the algorithm makes its predictions. As the number of trees increases, so does the accuracy of the predictions. In essence, Random Forest addresses the limitations of individual decision tree algorithms, enhancing accuracy and mitigating overfitting issues commonly associated with datasets [7].

The ability to handle datasets with a large number of predictor variables is a major advantage of employing random forest for prediction modelling; nevertheless, it is typically best to minimise the number of predictors needed to produce outcome predictions in order to enhance efficiency. For instance, when creating a medical prediction model, it may be preferable to include only a subset of the most significant factors rather than all of the variables present in the electronic medical record. Variable selection, which identifies the best predictors based on statistical parameters. Variable selection is frequently a crucial step in building prediction models, as many modern datasets contain hundreds or thousands of potential predictors [8].

2.4 Comparative Study

This section shows the existing research papers from other researchers that are similar to this research paper which each paper has its own result but has the same research goals. Table 1 below shows the comparison of previous research papers about machine learning for cardiovascular disease prediction.

Table 1 *Comparison Table*

Title of the research	Authors	Algorithm used	Description	Result
Heart disease prediction using machine learning algorithms	Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain, Preeti Nagrath	• Logistic regression • K-Nearest Neighbors (KNN) • Random Forest Classifier	This research is to check whether the patient is likely to be diagnosed with any cardiovascular heart disease based on their medical attributes such as gender, age, chest pain, fasting sugar level, etc.	Logistic Regression and KNN outperform Random Forest Classifier in the prediction of the patient diagnosed with heart disease [9].
Prediction of cardiovascular diseases based on machine learning	Weicheng Sun, Ping Zhang, Zilin Wang, Dongxu Li	• Support Vector Machine (SVM) • Logical regression • Random Forest	This research paper is to adopt machine learning-based methods including SVM, Logical regression, and Random forest to predict cardiovascular disease.	The performance of SVM is better than Logical regression and Random forest [10].

Table 1 *(cont.)*

Title of the research	Authors	Algorithm used	Description	Result
Machine learning prediction in cardiovascular diseases: a meta-analysis	Chayakrit Krittanawong, Hafeez Ul Hassan Virk, Sripal Bangalore, Zhen Wang, Kipp W. Johnson, Rachel Pinotti, HongJu Zhang, Scott Kaplin, Bharat Narasimhan, Takeshi Kitai, Usman Baber, Jonathan L. Halperin & W. H. Wilson Tang	• Convolutional Neural Network (CNN) • SVM • Boosting algorithm • Custom-built algorithm • Random Forest	This research paper aims to assess and summarize the overall predictive ability of ML algorithms in cardiovascular diseases.	SVM and boosting algorithms are widely used in cardiovascular medicine with good results [11].
Heart disease prediction using machine learning techniques : a survey	V.V. Ramalingam, Ayantan Dandapath, M Karthik Raja	• SVM • KNN • Naïve Bayes • Decision Tree • Random Forest • Ensemble Models	This research paper presents a survey of various models based on such algorithms and techniques and analyze their performance.	Each of the above-mentioned algorithms have performed extremely well in some cases but poorly in some other cases [12].
Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques	S. Mohan, C. Thirumalai and G. Srivastava	• Decision Tree • Language Model • SVM • Random Forest • Naïve Bayes • Neural Network • KNN • Generalized Linear Model (GLM) • Logistic Regression • Deep Learning • Gradient Boosted Tree • Hybrid Random Forest with Linear Model (HRFLM)	This research paper is to propose a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease.	The proposed hybrid HRFLM approach is used combining the characteristics of Random Forest and Linear Method [13].

Table 1 *(cont.)*

Title of the research	Authors	Algorithm used	Description	Result
Machine Learning Techniques For Heart Disease Prediction	A.Lakshmanarao, Y.Swathi, P.Sri Sai Sundareswar	• Logistic Regression • KNN • AdaBoost • Decision Tree • Naïve Bayes • Random Forest • SVM • Extra Tree Classifier • Gradient Boosting	This research paper uses machine learning techniques for the detection of heart disease	For random oversampling, SVM given the best accuracy. For Synthetic Minority Oversampling, Random Forest and Extra Tree Classifier given the best accuracy. For Adaptive synthetic sampling, Random Forest and Extra Tree Classifier given the best accuracy [14].
Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison	Md Mamun Ali, Bikash Kumar Paul, Kawsar Ahmed, Francis M. Bui, Julian M.W. Quinn, Mohammad Ali Moni	• KNN • Decision Tree • Random Forest • AdaboostM1 • Logistic Regression • Multilayer perception	This research study aimed to identify machine learning classifiers with the highest accuracy for heart disease diagnostic purposes.	Three classification algorithms KNN, Random Forest and Decision Tree performed extremely well with 100% accuracy [15].
A Cardiovascular Disease Prediction using Machine Learning Algorithms	Rubini PE, Dr.C.A.Subasini, Dr.A. Vanitha Katharine, V.Kumaresan, S.GowdhamKumar, T.M. Nithya	• Random Forest • Logistic Regression • SVM • Naïve Bayes	This research paper presents a comparative analysis of machine learning techniques like Random Forest, Logistic Regression, SVM, and Naïve Bayes in the classification of cardiovascular disease.	By the comparative analysis, machine learning algorithm Random Forest has proven to be the most accurate and reliable algorithm [16].

3. Methodology/Framework

In this chapter, the research framework will discuss the flow of research about the machine learning approach for cardiovascular disease prediction. A research framework holds significance as it outlines the complete flow, ensuring the research stays aligned with its stated objectives. This study will adhere to five defined phases, as depicted in **Fig. 3**. It aims to employ machine learning algorithms such as logistic regression and random forest for predicting cardiovascular disease. The primary goal anticipated will be the accuracy comparison between these two algorithms.

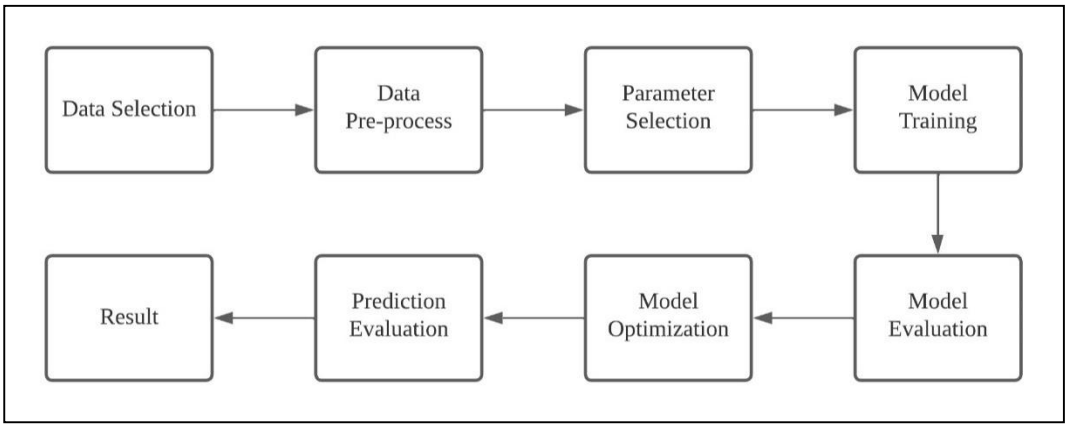


Fig. 3 *Research framework for cardiovascular disease prediction*

3.1 Data Selection

Datasets are essential in machine learning. Without the datasets, the further process of machine learning algorithms will be meaningless.

In 1988, the dataset for heart disease was released by UC Irvine Machine Learning Repository which has a collection of databases from Cleveland, Hungarian, Switzerland, Long Beach and Stalog for ML researchers to make empirical analyses of machine learning algorithms and predictions. Furthermore, this dataset was also released in a community of data scientists and machine learners known as Kaggle. To this day, many ML researchers still engage Kaggle datasets in the research literature for many projects. This shows the significance of datasets for reference in exploring analysis and predictive research. It plays an important role of summarization, pattern identification, and visualization in the data science field. Thus, it is the most suitable preferable of usage in predictive system research.

This research study uses two datasets from Kaggle. One of the datasets involved in this research study consists of 1190 data with 11 attributes and 1 target variable collected from patients [17]. Most of the attributes of the datasets are anonymized. The attributes that will be used in this research study are the age of the patient, sex of the patient, chest pain type, resting blood pressure, cholesterol, blood sugar levels on fasting, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, exercise-induced ST-depression in comparison with the state of rest (oldpeak), ST segment measured in terms of slope during peak exercise (ST slope) and target variable. Another dataset comprises 303 data with 13 attributes and 1 target variable [18]. The attributes are age of the patient, sex of the patient, exercise-induced angina (exng), chest pain type (cp), resting blood pressure (trtbps), cholestoral (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalachh), old exercise-induced angina (oldpeak), number of major vessels (caa), the slope of the peak exercise ST segment (slope), thallium 201 stress scintigraphy (thall) and target variable.

3.2 Data pre-processing

Data pre-processing is a crucial stage in the data analysis pipeline that involves cleaning, transforming, and organizing raw data before it's fed into a machine learning algorithm or statistical model for analysis [19]. The purpose of data pre-processing is to ensure that the data is in a format that enhances the performance and accuracy of the research model.

Data pre-processing involves various procedures, including handling missing values, addressing outliers, scaling or normalizing features, encoding categorical variables into numerical formats, selecting relevant features, and splitting the data into training, validation, and test sets. This process significantly influences the effectiveness and reliability of the insights or predictions derived from the data analysis.

In this research study, a function like .isna() is used to check for missing values in a DataFrame. Next .sum() is used to calculate the total number of missing values in each column. Furthermore, the .duplicated() method is used to check for duplicate rows in a DataFrame, and the .sum() method is used again to calculate the total number of duplicate rows. Moreover, the .drop_dupicates() method is used to remove duplicate rows from a DataFrame and followed by the inplace=True parameter to specify that the changes should be made to the original DataFrame, rather than creating a new DataFrame.

3.3 Parameter Selection

A parameter is a variable or attribute that is fixed for a particular function or algorithm. It is typically used to control the behavior of the function or algorithm. To apply the suggested methodology, several dataset parameters

Applied Information Technology and Computer Science Vol.6 No.1 (2025) p. 100-119

need consideration for the research experiment. The parameters have 2 categories; one is nominal, and the other one is numerical. The numerical parameters are parameters that involve numerical values and can be measured or quantified. For example, height, weight and temperature allow for mathematical operation. On the other hand, nominal parameters are categorical data with no inherent numerical value. They represent categories or names that don't have a quantitative value. In this study, the numerical parameters are age, resting blood pressure, cholesterol, maximum heart rate achieved and exercise-induced ST-depression in comparison with the state of rest, and number of major vessels. The nominal parameters are explained in Table 2 until Table 9 below.

Table 2 *Sex*

ID	Attribute
0	Female
1	Male

Table 3 *Chest pain type*

ID	Attribute
0	Typical Angina
1	Atypical Angina
2	Non-angina pain
3	Asymptomatic

Table 4 *Fasting blood sugar>120mg/dl*

ID	Attribute
1	True
0	False

Table 5 *Resting electrocardiographic results*

ID	Attribute
0	Normal
1	Having ST-T wave abnormality
2	Showing probable or definite left ventricular hypertrophy by Estes' criteria

Table 6 *Exercise induced angina*

ID	Attribute
0	No
1	Yes

Table 7 *The slope of the peak exercise ST segment*

ID	Attribute
0	Unsloping
1	Flat
2	Downsloping

Table 8 *Thallium 201 stress scintigraphy*

ID	Attribute
1	Normal
2	Fixed defect
3	Reversible defect

Table 9 *Target*

ID	Attribute
0	Patient is normal
1	Patient is suffering from heart risk

3.4 Model Training

Model training is a crucial step in which the selected machine learning algorithm learns patterns and relationships within the training dataset. The objective is to enable the model to make accurate predictions on new, unseen data. This process involves adjusting the internal parameters of the model based on the training data to minimize the difference between predicted outcomes and actual outcomes. Training dataset helps the machine learning algorithm to perform the correctional task, which helps to learn the attribute, classifying and clustering the dataset, while the testing dataset learned from the trained data and calculates its performance and accuracy can be predicted by the machine learning algorithm. The ratio used in this research is 7:3, in which 70% of the data is used in training data and 30% of the data is used for testing data.

3.5 Model Evaluation

Model evaluation is the process of determining the performance analysis of the dataset and machine learning algorithm using different metrics such as confusion matrix, precision, recall, F1 score, accuracy and receiver operating characteristic curve (ROC).

3.6 Model Optimization

Hyperparameter tuning is essential for optimizing machine learning models, and GridSearchCV is a popular method for this purpose. It systematically searches through a grid of specified parameters to find the best combination for a given model[20]. For logistic regression, key hyperparameters to tune include penalty, C, and solver. The penalty determines the type of regularization (l1, l2, or none) to prevent overfitting, while the C parameter controls the regularization strength, balancing model complexity and accuracy. The solver specifies the optimization algorithm with options like liblinear, lbfgs, and saga. For Random Forest models, important hyperparameters include n_estimators (number of trees), max_depth (maximum tree depth), min_samples_split (minimum samples to split a node), min_samples_leaf (minimum samples for a leaf node), and max_features (number of features considered for splits). GridSearchCV evaluates different combinations of these parameters to find the set that maximizes performance metrics like accuracy. This tuning process is crucial for enhancing the effectiveness of both Logistic Regression and Random Forest models in various machine learning applications.

3.7 Prediction Evaluation

In this research study, the prediction depends on the test set accuracy value to determine which algorithms are the best for the chosen datasets. There will be two accuracies where: one is the default accuracy, and the other one is the accuracy obtained after the hyperparameter tuning process. Default accuracy refers to the accuracy of a machine learning model using its initial default hyperparameter settings without explicitly specifying hyperparameters. Thus, giving a baseline performance of the model. On the other hand, Accuracy after hyperparameter tuning refers to the accuracy achieved by the model after optimizing its hyperparameters, which involves systematically searching for the best set of hyperparameters to maximize the model's performance on the datasets.

3.8 Hardware Requirement

The hardware used in this research study is shown in Table 10 below.

Table 10 *Hardware requirement*

Hardware	Description
ASUS TUF Gaming F15	• Intel(R) Core(TM) i5-10300H CPU @ 2.50GHz 2.50 GHz • Installed Memory (RAM): 16GB • System Type: 64-bit operating system, x64-based processor • Operating System: Windows 10

Applied Information Technology and Computer Science Vol.6 No.1 (2025) p. 100-119

3.9 Software Requirement

The data analysis tool used in this research study is called Google Colab. Machine learning projects benefit significantly from the use of Google Colab. Google Colab offers remarkable advantages as a free, accessible tool. It operates entirely in the cloud, eliminating the need for downloads or subscriptions, and is able to operate with just using a Google account. Google Colab offers advanced functionalities including support for GPUs and TPUs, which prove immensely beneficial when executing machine learning and deep learning models. Moreover, it is also pre-installed with an array of libraries like NumPy, pandas, and matplotlib, streamlining access to their functionalities. Additionally, users can also import and utilize machine-learning libraries. For example, Sklearn, TensorFlow, and PyTorch have been used to create custom machine learning models [21].

4. Result and Analysis

This chapter will present the result of the analysed dataset in tables and graph forms. Two datasets are employed: one containing 14 attributes and the other 12 attributes. These datasets are analysed separately to compare the resulting predictive performance. Each dataset is split into training and testing subsets during the experimentation phase. The Random Forest and Logistic Regression algorithms are applied to each dataset. Finally, an analysis of the results will be presented.

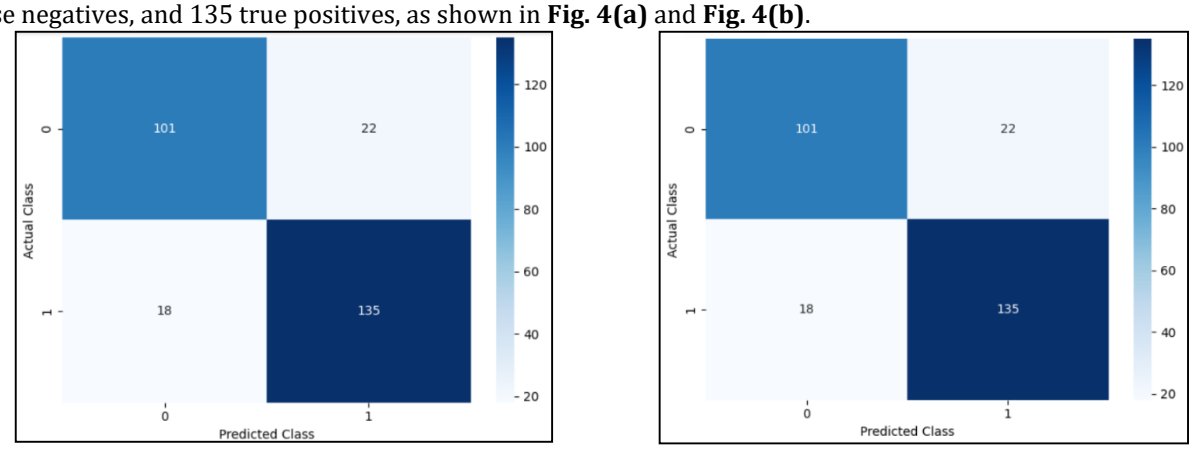
4.1 Experiment Phase

In the experimental phase, classifier performance evaluation is performed on Google Colab, a Python-based platform for developing predictive models. The algorithms under consideration, Logistic Regression and Random Forest, are assessed based on their accuracy, precision, recall, F1-score, and confusion matrix when analyzing the datasets. The data is divided into two datasets for training and testing, with a 70:30 split ratio applied randomly during the experiment. Consequently, 70% of the data is used for training, while the remaining 30% is used for testing.

4.2 Result of Performance Metrics

4.2.1 Logistic Regression Result For Dataset A

In this study, the performance of logistic regression models for predicting cardiovascular disease using Dataset A was evaluated. A default logistic regression model with an optimized model incorporating hyperparameters is compared. The confusion matrices for both models were identical, with 101 true negatives, 22 false positives, 18 false negatives, and 135 true positives, as shown in **Fig. 4(a)** and **Fig. 4(b)**.



(a) **(b)**

Fig 4 *Confusion Matrix (a)Default Model; (b)Best Model*

The classification reports for both models were also the same, as shown in **Fig. 5(a)** and **Fig. 5(b)**. For class 0, the precision was 0.85, the recall was 0.82, and the F1-score was 0.83, with 123 instances. For class 1, the precision was 0.86, the recall was 0.88, and the F1-score was 0.87, with 153 instances. These metrics resulted in a macro average of 0.85 for precision, recall, and F1-score, and a weighted average of 0.85 for precision and F1-score, and 0.86 for recall.

Classification Report for the test set (Default Model):

	precision	recall	f1-score	support
0	0.85	0.82	0.83	123
1	0.86	0.88	0.87	153
accuracy			0.86	276
macro avg	0.85	0.85	0.85	276
weighted avg	0.85	0.86	0.85	276

(a)

Classification Report for the test set (Model with Hyperparameters):

	precision	recall	f1-score	support
0	0.85	0.82	0.83	123
1	0.86	0.88	0.87	153
accuracy			0.86	276
macro avg	0.85	0.85	0.85	276
weighted avg	0.85	0.86	0.85	276

(b)

Fig. 5 *Classification Report (a) Default Model; (b) Best Model*

Fig. 6 below shows the feature importance analysis revealed that the most influential features in predicting cardiovascular disease were ST slope, chest pain type, and sex. These features played a significant role in the model's decision-making process, highlighting their importance in the context of cardiovascular disease prediction.

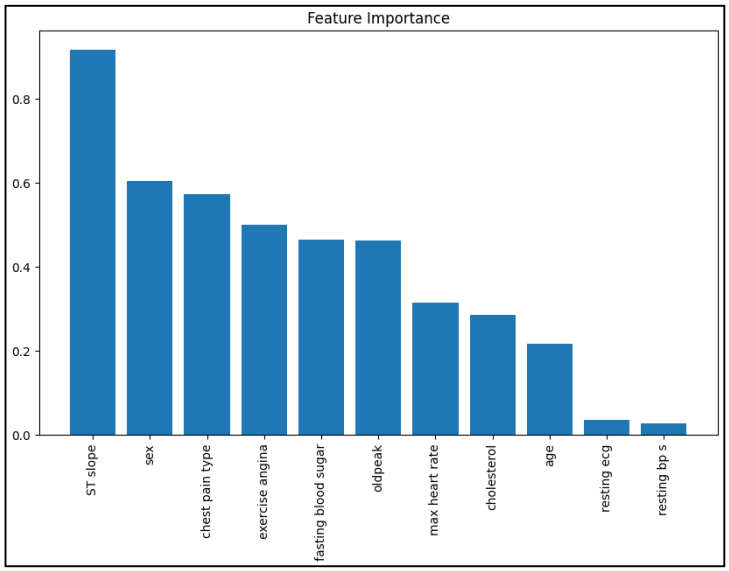
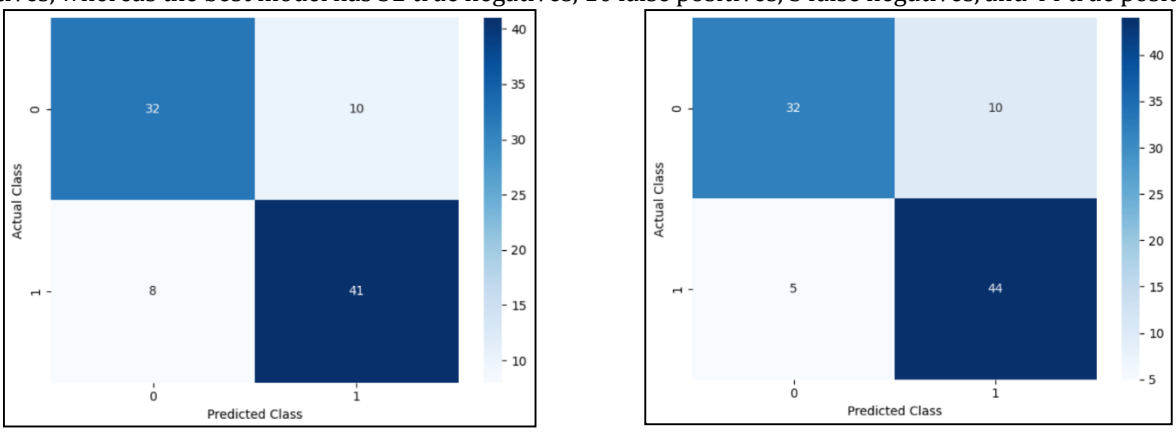


Fig. 6 *Feature Importance of the Dataset A for Logistic Regression*

4.2.2 Logistic Regression Result For Dataset B

The comparison between the default and best logistic regression models for cardiovascular disease prediction on Dataset B reveals insightful findings. The confusion matrices for both models are similar, shown in **Fig. 7(a)** and **Fig. 7(b)**, with the default model showing 32 true negatives, 10 false positives, 8 false negatives, and 41 true positives, whereas the best model has 32 true negatives, 10 false positives, 5 false negatives, and 44 true positives.



(a) **(b)**

Fig. 7 *Confusion Matrix (a)Default Model; (b)Best Model*

The classification report in **Fig. 8(a)** provides a detailed breakdown of precision, recall, and F1-score for each class. For the default model, the precision for class 0 is 0.80, indicating that 80% of the predicted class 0 instances

are correct, and for class 1, it is 0.80, meaning that 80% of the predicted class 1 instances are correct. Recall for class 0 is 0.76, indicating that 76% of actual class 0 instances are correctly identified, and for class 1, it is 0.84, showing that 84% of actual class 1 instances are correctly identified. The F1-score, which is the harmonic mean of precision and recall, is 0.78 for class 0 and 0.82 for class 1. The overall accuracy of the default model is 0.80, with macro and weighted averages for precision, recall, and F1-score all at 0.80.

For the best model, after hyperparameter tuning shown in **Fig. 8(b)**, the precision for class 0 is 0.86, while for class 1, it is 0.81. Recall for class 0 remains at 0.76, and for class 1 is 0.90. The F1-score is 0.81 for class 0 and 0.85 for class 1. The overall accuracy of the best model is 0.84, with macro and weighted averages for precision, recall, and F1-score all at 0.84. Despite the marginally lower precision for class 1 compared to the default model, the best model demonstrates improved recall for class 1 and higher overall accuracy, indicating a more balanced and reliable performance due to the thorough hyperparameter tuning process.

Classification Report for the test set (Default Model):

	precision	recall	f1-score	support
0	0.80	0.76	0.78	42
1	0.80	0.84	0.82	49
accuracy			0.80	91
macro avg	0.80	0.80	0.80	91
weighted avg	0.80	0.80	0.80	91

(a)

Classification Report for the test set (Model with Hyperparameters):

	precision	recall	f1-score	support
0	0.86	0.76	0.81	42
1	0.81	0.90	0.85	49
accuracy			0.84	91
macro avg	0.84	0.83	0.83	91
weighted avg	0.84	0.84	0.83	91

(b)

Fig. 8 *Classification Report (a) Default Model; (b) Best Model*

Fig. 9 shows the feature importance analysis from the best model highlights chest pain type (cp), exercise-induced angina (exng), and sex as the most significant predictors of cardiovascular disease.

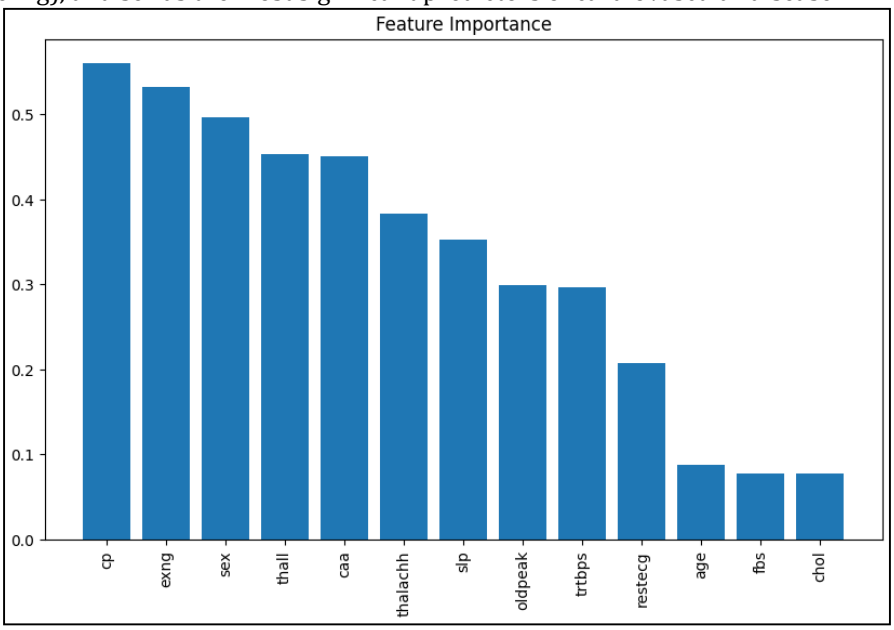
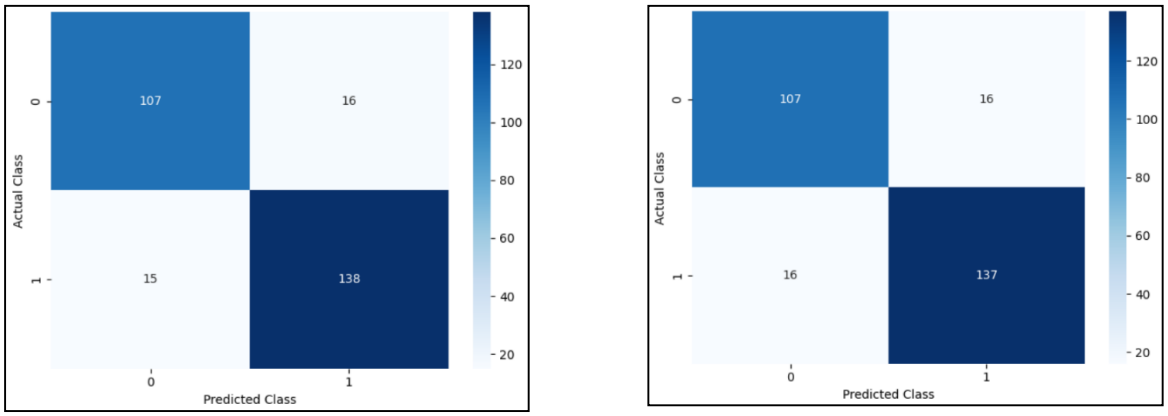


Fig. 9 *Feature Importance of the Dataset B for Logistic Regression*

4.2.3 Random Forest Result For Dataset A

The comparison between the default and best Random Forest models for cardiovascular disease prediction on Dataset A reveals insightful findings. The confusion matrices for both models are quite similar as shown in **Fig. 10(a)** and **Fig. 10(b)**, with the default model showing 107 true negatives, 16 false positives, 15 false negatives, and 138 true positives, whereas the best model has 107 true negatives, 16 false positives, 16 false negatives, and 137 true positives.



(a) **(b)**

Fig. 10 *Confusion Matrix (a)Default Model; (b)Best Model*

The classification report in **Fig. 11(a)** provides a detailed breakdown of precision, recall, and F1-score for each class. For the default model, the precision for class 0 is 0.88, indicating that 88% of the predicted class 0 instances are correct, and for class 1, it is 0.90, meaning that 90% of the predicted class 1 instances are correct. Recall for class 0 is 0.87, indicating that 87% of actual class 0 instances are correctly identified, and for class 1, it is 0.90, showing that 90% of actual class 1 instances are correctly identified. The F1-score, which is the harmonic mean of precision and recall, is 0.87 for class 0 and 0.90 for class 1. The overall accuracy of the default model is 0.89, with macro and weighted averages for precision, recall, and F1-score all at 0.89.

For the best model, after hyperparameter tuning shown in **Fig. 11(b)**, the precision for class 0 is slightly lower at 0.87, while for class 1, it remains at 0.90. Recall for class 0 is 0.87, and for class 1 is 0.90, the same as the default model. The F1-score is also 0.87 for class 0 and 0.90 for class 1. The overall accuracy of the best model is 0.88, with macro and weighted averages for precision, recall, and F1-score all at 0.88. Despite the marginally lower accuracy and precision for class 0, the best model is more robust and reliable due to the thorough hyperparameter tuning process, ensuring better generalizability.

Classification Report for the test set (Default Model):

	precision	recall	f1-score	support
0	0.88	0.87	0.87	123
1	0.90	0.90	0.90	153
accuracy			0.89	276
macro avg	0.89	0.89	0.89	276
weighted avg	0.89	0.89	0.89	276

(a)

Classification Report:

	precision	recall	f1-score	support
0	0.87	0.87	0.87	123
1	0.90	0.90	0.90	153
accuracy			0.88	276
macro avg	0.88	0.88	0.88	276
weighted avg	0.88	0.88	0.88	276

(b)

Fig. 11 *Classification Report (a) Default Model; (b) Best Model*

Fig. 12 shows the feature importance analysis from the best model highlights ST slope, chest pain type, and old peak as the most significant predictors of cardiovascular disease.

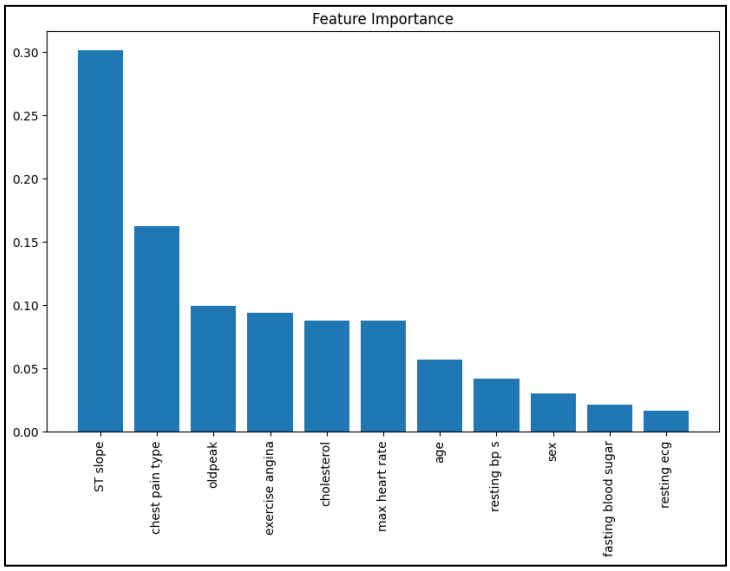
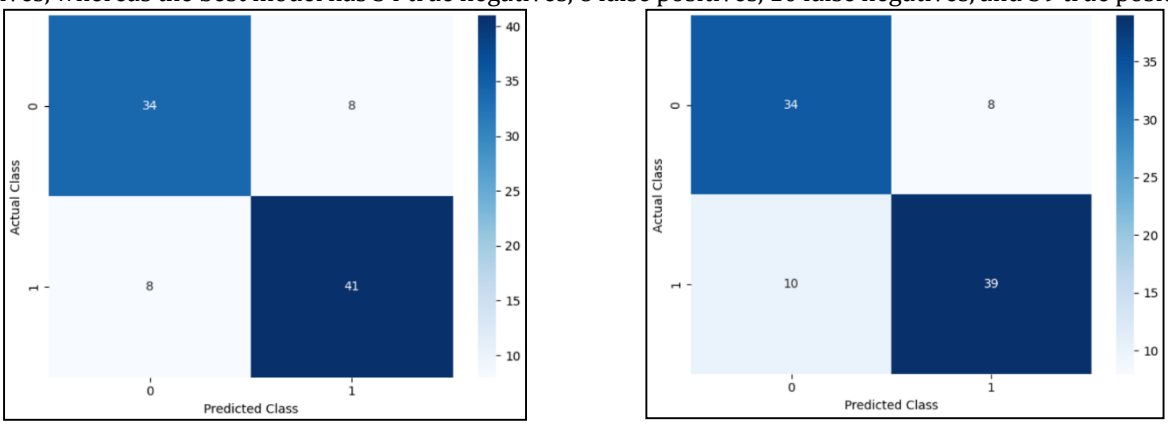


Fig.12 *Feature Importance of the Dataset A for Random Forest*

4.2.4 Random Forest Result For Dataset B

The comparison between the default and best Random Forest models for cardiovascular disease prediction on Dataset B reveals insightful findings. The confusion matrices for both models are similar, shown in **Fig. 13(a)** and **Fig. 13(b)**, with the default model showing 34 true negatives, 8 false positives, 8 false negatives, and 41 true positives, whereas the best model has 34 true negatives, 8 false positives, 10 false negatives, and 39 true positives.



(a) **(b)**

Fig. 13 *Confusion Matrix (a)Default Model; (b)Best Model*

The classification report in **Fig.14(a)** provides a detailed breakdown of precision, recall, and F1-score for each class. For the default model, the precision for class 0 is 0.81, indicating that 81% of the predicted class 0 instances are correct, and for class 1, it is 0.84, meaning that 84% of the predicted class 1 instances are correct. Recall for class 0 is 0.81, indicating that 81% of actual class 0 instances are correctly identified, and for class 1, it is 0.84, showing that 84% of actual class 1 instances are correctly identified. The F1-score, which is the harmonic mean of precision and recall, is 0.81 for class 0 and 0.84 for class 1. The overall accuracy of the default model is 0.82, with macro and weighted averages for precision, recall, and F1-score all at 0.82.

For the best model, after hyperparameter tuning shown in **Fig. 14(b)**, the precision for class 0 is slightly lower at 0.77, while for class 1, it is 0.83. Recall for class 0 is 0.81, and for class 1 is 0.80. The F1-score is 0.79 for class 0 and 0.81 for class 1. The overall accuracy of the best model is 0.80, with macro and weighted averages for precision, recall, and F1-score all at 0.80. Despite the marginally lower accuracy and precision for class 0, the best model is more robust and reliable due to the thorough hyperparameter tuning process, ensuring better generalizability.

Classification Report for the test set (Default Model):

	precision	recall	f1-score	support
0	0.81	0.81	0.81	42
1	0.84	0.84	0.84	49
accuracy			0.82	91
macro avg	0.82	0.82	0.82	91
weighted avg	0.82	0.82	0.82	91

(a)

Classification Report:

	precision	recall	f1-score	support
0	0.77	0.81	0.79	42
1	0.83	0.80	0.81	49
accuracy			0.80	91
macro avg	0.80	0.80	0.80	91
weighted avg	0.80	0.80	0.80	91

(b)

Fig. 14 *Classification Report (a) Default Model; (b) Best Model*

Fig. 15 shows the feature importance analysis from the best model highlights oldpeak, chest pain type (CP), and thallium 201 stress scintigraphy (thall) as the most significant predictors of cardiovascular disease.

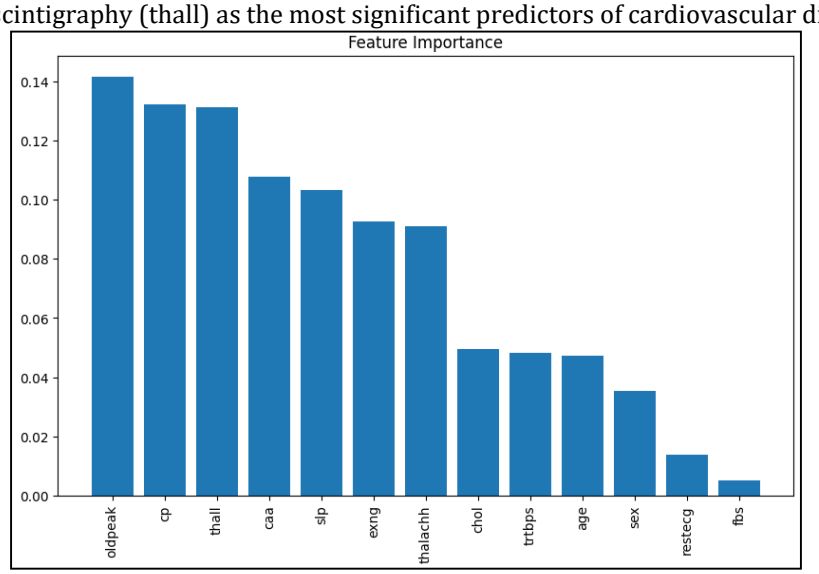


Fig. 15 *Feature Importance of the Dataset B for Random Forest*

4.3 Accuracy of the Algorithms

4.3.1 Default Model

Through the experimental phase, the accuracy shown in the default model for Logistic Regression is 85.5072% and 80.2198%, while for Random Forest, it is 88.7681% and 82.4176% as shown in **Fig. 16**.

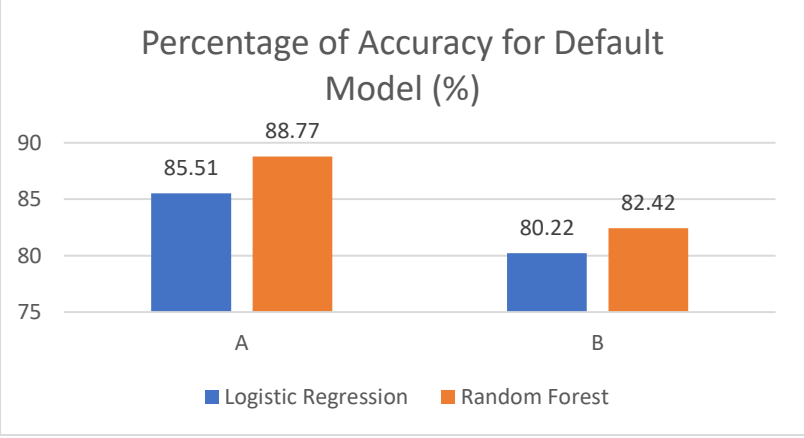


Fig. 16 *Visualization Graph for Default Model*

4.3.2 After Hyperparameter Tuning (Best Model)

Through the experimental phase, the accuracy shown in the best model for Logistic Regression is 85.5072% and 83.5165%, while for Random Forest, it is 88.4058% and 80.2198%, as shown in **Fig. 17**.

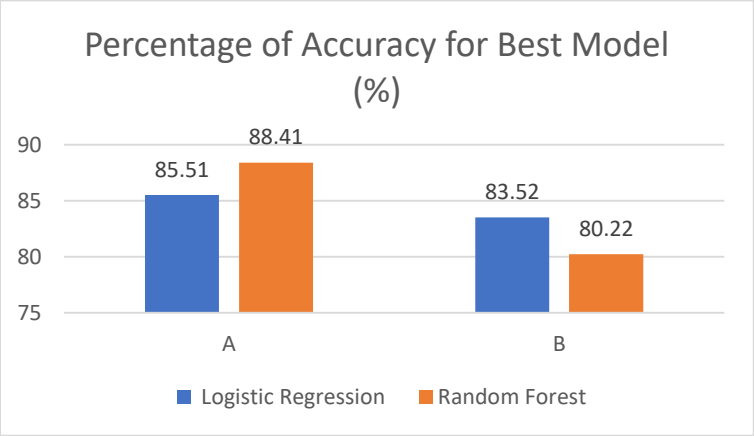


Fig. 17 *Visualization Graph for Best Model*

4.4 ROC Curve for the Algorithms

4.4.1 ROC Curve for Logistic Regression in Dataset A

The ROC curve shown in **Fig. 18** is for a logistic regression model applied to Dataset A. It demonstrates the model's performance in distinguishing between positive and negative classes. The x-axis represents the False Positive Rate (FPR), and the y-axis represents the True Positive Rate (TPR). The curve deviates significantly from the diagonal line (which indicates a random classifier) and stays close to the top left corner, indicating high model performance. The Area Under the Curve (AUC) is 0.92, which confirms that the logistic regression model has a high level of accuracy in classification tasks, effectively distinguishing between positive and negative instances.

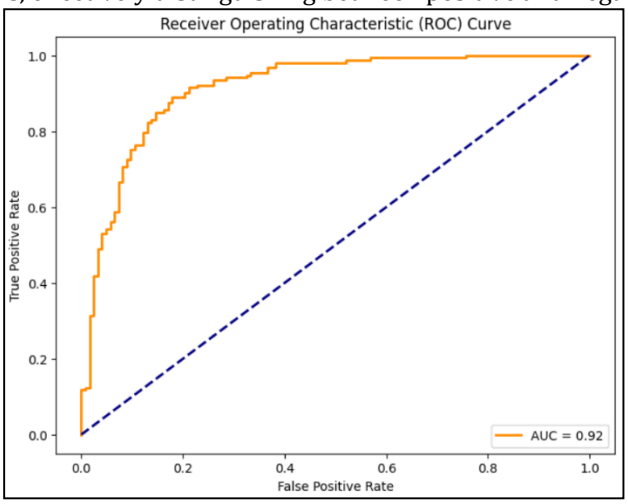


Fig. 18 *ROC Curve for Logistic Regression in Dataset A*

4.4.2 ROC Curve for Logistic Regression in Dataset B

The ROC curve shown in **Fig. 19** is for a logistic regression model applied to Dataset B. The x-axis represents the False Positive Rate (FPR), and the y-axis represents the True Positive Rate (TPR). The curve deviates from the diagonal line, which represents the performance of a random classifier, and stays relatively close to the top left corner, indicating good model performance. The Area Under the Curve (AUC) is 0.89, which suggests that the logistic regression model has a high level of accuracy in distinguishing between positive and negative classes in Dataset B. This AUC value indicates that the model has strong discriminative capability, though slightly less than the previous model in dataset A, with an AUC of 0.92.

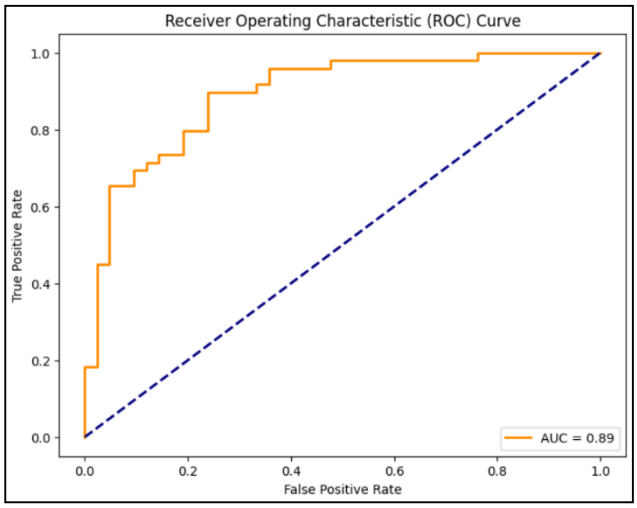


Fig. 19 *ROC Curve for Logistic Regression in Dataset B*

4.4.3 ROC Curve for Random Forest in Dataset A

The ROC curve shown in **Fig. 20** is for a Random Forest model applied to Dataset A. The x-axis represents the False Positive Rate (FPR), while the y-axis represents the True Positive Rate (TPR). The curve stays close to the top left corner and deviates significantly from the diagonal line, representing a random classifier's performance. This indicates excellent model performance. The Area Under the Curve (AUC) is 0.94, suggesting that the Random Forest model has a very high level of accuracy in distinguishing between positive and negative classes in Dataset A. This high AUC value indicates the model's strong discriminative capability.

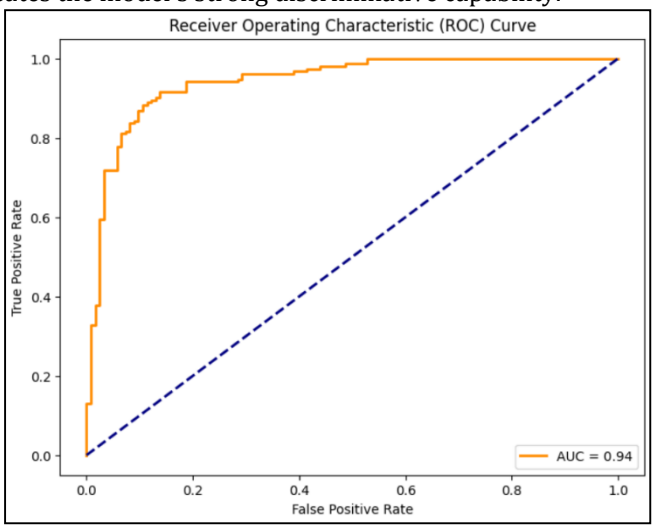


Fig. 20 *ROC Curve for Random Forest in Dataset A*

4.4.4 ROC Curve for Random Forest in Dataset B

The ROC curve shown in **Fig. 21** is for a Random Forest model applied to Dataset B. The x-axis represents the False Positive Rate (FPR), while the y-axis represents the True Positive Rate (TPR). The curve stays close to the top left corner and deviates significantly from the diagonal line, representing a random classifier's performance. This indicates excellent model performance. The Area Under the Curve (AUC) is 0.86, suggesting that the Random Forest model has a very high level of accuracy in distinguishing between positive and negative classes in Dataset B. This AUC value indicates that the model has a very strong discriminative capability, though slightly less than the previous model in dataset A with an AUC of 0.94.

Applied Information Technology and Computer Science Vol.6 No.1 (2025) p. 100-119

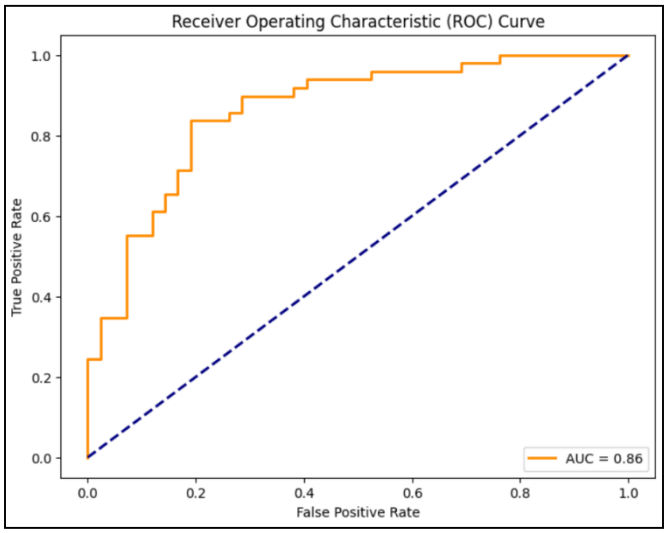


Fig. 21 *ROC Curve for Random Forest in Dataset B*

4.5 Result Summary

For Dataset A, two algorithms were tested: Logistic Regression and Random Forest. Logistic Regression achieved an accuracy of 85.51% with a ROC AUC of 0.92, and its confusion matrix showed 101 true negatives, 22 false positives, 18 false negatives, and 135 true positives. The most important features in the Logistic Regression model were ST slope, chest pain type, and sex. On the other hand, the Random Forest model outperformed Logistic Regression with an accuracy of 88.77% and a ROC AUC of 0.94. Its confusion matrix indicated 107 true negatives, 16 false positives, 15 false negatives, and 138 true positives. After hyperparameter tuning, the Random Forest model's accuracy slightly decreased to 88.41%, with its confusion matrix showing 107 true negatives, 16 false positives, 16 false negatives, and 137 true positives. The key features identified in the Random Forest model were ST slope, chest pain type, and old peak. In conclusion, Random Forest is more suitable for Dataset A due to its higher accuracy and ROC AUC, with ST slope being the most significant feature in both models. Additionally, Random Forest had a higher number of true positives and true negatives, making it more accurate in classifying both classes.

For Dataset B, Logistic Regression and Random Forest were also evaluated. The default Logistic Regression model achieved an accuracy of 80.22% and a ROC AUC of 0.89, with a confusion matrix of 34 true negatives, 8 false positives, 8 false negatives, and 41 true positives. Key features for Logistic Regression included chest pain type (cp), exercise-induced angina (exng), and sex. After hyperparameter tuning, the accuracy of Logistic Regression improved to 83.52%, and its confusion matrix indicated 36 true negatives, 6 false positives, 9 false negatives, and 40 true positives. The Random Forest model had an accuracy of 82.42% and a ROC AUC of 0.86, with its confusion matrix showing 34 true negatives, 8 false positives, 8 false negatives, and 41 true positives. After tuning, the Random Forest model's accuracy was 80.22%, with a confusion matrix of 34 true negatives, 8 false positives, 10 false negatives, and 39 true positives. The significant features for Random Forest were old peak, chest pain type (cp), and thallium 201 stress scintigraphy (thall). In conclusion, Logistic Regression is more suitable for Dataset B after hyperparameter tuning due to its higher accuracy and ROC AUC compared to Random Forest. The most important features for Logistic Regression were chest pain type, exercise-induced angina, and sex. Logistic regression also showed better performance with higher true negatives and maintained a substantial number of true positives, ensuring more accurate classifications of both classes.

5. Conclusion

This research comprehensively evaluated the performance of Logistic Regression and Random Forest algorithms on two distinct datasets, Dataset A and Dataset B, specifically for predicting cardiovascular disease. The findings indicated that Random Forest outperformed Logistic Regression on Dataset A, achieving superior accuracy and ROC AUC values. Notably, the feature importance analysis identified ST slope, chest pain type, and old peak as the most significant predictors for the Random Forest model. In contrast, Logistic Regression demonstrated better performance on Dataset B, especially after hyperparameter tuning, achieving higher accuracy and ROC AUC than Random Forest. For Logistic Regression, key predictive features included chest pain type, exercise-induced angina, and sex, which substantially enhanced the model's predictive power. This study underscores the critical importance of selecting the right machine learning algorithm based on the specific characteristics of the dataset, and the detailed feature importance analysis provided deeper insights into the variables that most significantly influenced model outcomes. The research contributes valuable knowledge to the field of machine learning,

particularly in the context of cardiovascular disease prediction. Future work should focus on expanding the size and diversity of the datasets to enhance generalizability, exploring additional machine learning algorithms, employing advanced feature engineering techniques, and validating the models in real-world scenarios to ensure their practical utility and robustness.

Acknowledgement

The authors would like to thank the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, for its support.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *Low Yin Qian, Dr. Noor Zuraidin bin Mohd Safar;* ***data collection:*** *Low Yin Qian, Dr. Noor Zuraidin bin Mohd Safar;* ***analysis and interpretation of results:*** *Low Yin Qian, Dr. Noor Zuraidin bin Mohd Safar;* ***draft manuscript preparation:*** *Low Yin Qian, Dr. Noor Zuraidin bin Mohd Safar. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] American Heart Association, "What is Cardiovascular Disease?"
[2] World Health Organization, "Cardiovascular diseases (CVDs)."
[3] C. Matt, "What is Machine Learning? Definition, Types, Tools & More." [Online]. Available: https://www.datacamp.com/blog/what-is-machine-learning
[4] J. Alzubi, A. Nayyar, and A. Kumar, "Machine Learning from Theory to Algorithms: An Overview," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Nov. 2018. doi: 10.1088/1742-6596/1142/1/012012.
[5] A. G, B. Ganesh, A. Ganesh, C. Srinivas, Dhanraj, and K. Mensinkal, "Logistic regression technique for prediction of cardiovascular disease," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 127–130, Jun. 2022, doi: 10.1016/j.gltp.2022.04.008.
[6] R. Bhuvana, S. Maheshwari, and S. Sasikala, "Predict the Heart Disease Using a Logistic Regression Classifier Algorithm," in *Proceedings of the 2023 12th International Conference on System Modeling and Advancement in Research Trends, SMART 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 649–652. doi: 10.1109/SMART59791.2023.10428486.
[7] M. Onesmus, "Introduction to Random Forest in Machine Learning." [Online]. Available: https://www.webscale.com/engineering-education/introduction-to-random-forest-in-machine-learning/
[8] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Systems with Applications*, vol. 134. Elsevier Ltd, pp. 93–101, Nov. 15, 2019. doi: 10.1016/j.eswa.2019.05.028.
[9] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing Ltd, Jan. 2021. doi: 10.1088/1757-899X/1022/1/012072.
[10] W. Sun, P. Zhang, Z. Wang, and D. Li, "Prediction of Cardiovascular Diseases based on Machine Learning," *ASP Transactions on Internet of Things*, vol. 1, no. 1, pp. 30–35, May 2021, doi: 10.52810/tiot.2021.100035.
[11] C. Krittanawong *et al.*, "Machine learning prediction in cardiovascular diseases: a meta-analysis," *Sci Rep*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-72685-1.
[12] V. V. Ramalingam, A. Dandapath, and M. Karthik Raja, "Heart disease prediction using machine learning techniques: A survey," *International Journal of Engineering and Technology(UAE)*, vol. 7, no. 2.8 Special Issue 8, pp. 684–687, 2018, doi: 10.14419/ijet.v7i2.8.10557.
[13] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
[14] S. Pullela, A. Lakshmanarao, Y. Swathi, P. Sri, and S. Sundareswar, "Machine Learning Techniques For Heart Disease Prediction Article in," *International Journal of Scientific & Technology Research*, vol. 8, no. 11, 2020, [Online]. Available: www.ijstr.org

[15] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Comput Biol Med*, vol. 136, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104672.
[16] R. Pe, C. A. Subasini, A. V. Katharine, V. Kumaresan, S. Gowdhamkumar, and T. M. Nithya, "A Cardiovascular Disease Prediction using Machine Learning Algorithms," 2021. [Online]. Available: http://annalsofrscb.ro
[17] S. Manu, "Heart Disease Dataset (Comprehensive)." [Online]. Available: https://www.kaggle.com/datasets/sid321axn/heart-statlog-cleveland-hungary-final/data
[18] R. RASHIK, "Heart Attack Analysis & Prediction Dataset." [Online]. Available: https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset
[19] G. Kechit, "Data Preprocessing in Machine Learning: 7 Easy Steps To Follow." [Online]. Available: https://www.upgrad.com/blog/data-preprocessing-in-machine-learning/
[20] "GridSearchCV," scikit-learn 1.5.0. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html#sklearn.model_selection.GridSearchCV
[21] S. Khaleel, "Introduction to Google Colab," LinkedIn. [Online]. Available: https://www.linkedin.com/pulse/introduction-google-colab-khaleel-shaik/