Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 29 (2026), No. 7 & 8, pp. 787–802
DOI : 10.47974/JSMS-1587

Heart disease prediction using logistic regression

S. Baby Latha *
R. Abinaya †
Department of Actuarial Science
Bishop Heber College (Affiliated to Bharathidasan University)
Trichy 620017
Tamil Nadu
India

Abstract

Heart disease remains a major global health concern and one of the leading causes of mortality, emphasizing the importance of early detection and reliable risk assessment. This research investigates the effectiveness of Logistic Regression (LR) in predicting heart disease risk using the Heart Disease Statlog dataset. Various clinical and demographic parameters including age, cholesterol levels, exercise-induced angina, and ST depression were examined to build a predictive model. The dataset was subjected to preprocessing, feature selection, and exploratory data analysis (EDA) before model training. Evaluation metrics such as accuracy (92.59%), precision (94.7%), recall (85.7%), and AUC-ROC (0.95) demonstrated strong model performance. The study underscores the interpretability of LR for medical diagnosis while recognizing its limitations in modeling complex non-linear patterns. Future studies may enhance prediction accuracy by employing advanced machine learning techniques and integrating real-time health monitoring systems. Overall, the findings support improvements in early disease detection, healthcare decision-making, and actuarial risk analysis in the health insurance sector.

Subject Classification: *15A06, 60G57, 62H12, 62H35.*

Keywords: *Heart disease, Predictive model, Logistic regression, Risk assessment, Machine learning.*

I. Introduction

According to the World Health Organization (WHO), cardiovascular diseases (CVDs) are the leading cause of death worldwide, responsible for approximately 17.9 million fatalities each year. These diseases include a

* *E-mail:* **babylatha.as@bhc.edu.in (Corresponding Author)**
† *E-mail:* **abinayarajagopal2@gmail.com**

wide spectrum of heart and blood vessel disorders, such as coronary artery disease, heart attacks, and arrhythmias. Lifestyle factors such as poor dietary habits, a lack of exercise, smoking, and excessive stress are major causative factors of the rising burden of heart disease. Other conditions, such as hypertension, diabetes, obesity, and hypercholesterolemia, also put a person at high risk of heart disease. Early diagnosis and prevention are the cornerstones of managing the occurrence of heart disease. Traditional diagnosis relies on clinical assessment, laboratory tests, and imaging studies. Advances in data analysis and machine learning have opened new avenues for statistical model-based heart disease prediction. Among these techniques, logistic regression has been put forward as an inexpensive yet powerful tool for classifying patients as high-risk or low-risk cases based on primary health predictors. Logistic regression is particularly suitable for this purpose because it is a technique of modelling the probability of an event (e.g., a diagnosis of heart disease) from a set of independent variables and is extremely interpretive and efficient for decision-making in the healthcare sectors.

From an actuarial science perspective, the ability to forecast heart disease accurately has significant actuarial implications for health insurance pricing and risk assessment. Insurers rely on forecasting models to estimate the health risk of individuals and determine appropriate premium levels. A correct model can allow insurers to structure risk-based pricing policies, keeping fair and sustainable premium levels. Predictive analytics may be applied to create preventive health programs, encouraging policyholders to adopt healthy lifestyles in order to reduce their insurance premiums. With the significance of early diagnosis, this research targets the creation of a logistic regression model to predict heart disease based on major clinical and demographic characteristics. Through the examination of real-world data, this study seeks to determine the most critical risk factors and the performance of logistic regression as a predictive model. The results of this research may help enhance early diagnosis, better risk assessment models, and more accurate actuarial pricing models in health insurance.

Cardiovascular disease is a huge burden on patients and healthcare systems worldwide. Even in the midst of swift changes in medical technology, unwarranted delay in diagnosis time and human errors in symptom interpretation can lead to suboptimal results. A comprehensive predictive model for heart disease may address two main challenges are Early Diagnosis and Treatment: Heart disease risk prediction based on clinical and lifestyle factors will enable the physicians to treat it at an early stage, thereby saving lives and reducing the cost of treatment and Reduction of Human Error: Predictive models can be utilized as clinician decision-support tools, complementing their expertise and reducing diagnostic inaccuracies. In addition to its health burden, heart disease also imposes a significant economic cost on patients, healthcare facilities, and insurance companies. Cardiovascular diseases, for example, have been projected by Bloom et al. (2021) to cost the healthcare system almost $1 trillion annually, in terms of hospitalization, medication, surgery, and lost productivity. Indirect costs, such as loss of the workforce and permanent disability, also affect economies, particularly in the elderly. Heart disease trends have an important influence on the insurance industry. Actuarial models rely on risk factor assessments to estimate premium rates and chances of claims. An effective heart disease forecasting model would enhance underwriting, allowing insurers to issue more individualized and equitable policies. Moreover, insurers are marketing more actively preventive health rewards, such as lower premiums for physical exercise or for joining smoking cessation programs.

II. Literature review

The World Health Organization (WHO, 2023) reports that cardiovascular diseases (CVDs)such as coronary artery disease (CAD), heart failure, and irregular heart rhythms account for about 32% of deaths globally, positioning them as the primary cause of mortality across the world. Of these, ischemic heart disease (IHD) and stroke are the most fatal. The Global Burden of Disease (GBD) study highlights that low- and middle-income nations have the highest burden of CVDs, and the cases increase with urbanization, dietary changes, and increasingly inactive lifestyles (Roth et al., 2020). The increased incidence of risk factors like obesity, hypertension, diabetes, and smoking further aggravates the burden of CVDs, and prevention and early detection are therefore imperative in decreasing mortality rates.

A wealth of research has proved that early medical intervention is necessary in reducing mortality from heart disease. Benjamin et al. [13], in the Journal of the American Heart Association (JAHA), highlighted the necessity of lifestyle modification dietetic modification, exercise, and cessation of smoking and the use of advanced risk prediction models in reducing the occurrence and severity of heart disease. Such prediction models enable doctors to diagnose patients with greater risk more effectively, which translates into early medical intervention and customized treatment strategies. Data-driven techniques, including statistical modeling and machine learning, have greatly improved the accuracy of prediction of heart disease risk. Such techniques offer a more objective data-driven strategy in the identification of high-risk individuals in addition to conventional diagnostic methods.

Its effectiveness in the field has been proven in various studies. Hossain et al. [2], achieved 90% accuracy with RF on Bangladeshi clinical data, proving its reliability across various population groups. Torthi et al. [1], introduced a hybrid RF model, which utilized Bat Algorithm and Particle Swarm Optimization, and outperformed traditional methods by achieving 98.71% accuracy through the UCI dataset. Lutimath, Sharma, and Byregowda [8], successfully deployed RF on the UCI Cleveland dataset, proving its capability to work on sophisticated biomedical data. Similarly, Yang et al. [9], deployed RF on a large Chinese dataset for the prediction of 3-year cardiovascular disease (CVD) risk, proving its capability to be scaled up for high-dimensional medical data.

Sabarinathan and Sugumaran [11], applied Decision Tree and SVM classifiers to diagnose heart disease using patient health records, achieving an accuracy of 84.2% and demonstrating the high interpretability of Decision Trees for clinical decision support. Akila and Chandramathi [10], developed a hybrid model combining Decision Tree with Multi-Layer Perceptron (MLP), improving predictive accuracy by 7% compared to standalone Decision Trees by leveraging the strengths of both techniques. Deepika [3], conducted a comparative analysis of various Decision Tree algorithms, including ID3, C4.5, and CART, and found that C4.5 performed best with an accuracy of 86.5%, showcasing its ability to handle categorical and continuous attributes effectively.

Narayan and Gobal [5], proposed an optimized Decision Tree-based fuzzy rule classifier enhanced by the Improved Cuckoo Search Algorithm, achieving 88.9% accuracy and demonstrating improved stability in handling imbalanced datasets. Pandey et al. [4] developed a heart disease prediction model based on Decision Tree algorithms utilizing clinical data from the Cleveland dataset, highlighting the crucial role of selecting relevant features. Their model identified blood pressure, cholesterol, and smoking status as key determinants in predicting heart disease risk and attained an accuracy of 85.3%. Support Vector Machines (SVMs) are widely applied in heart disease prediction due to their effectiveness in managing high-dimensional datasets and delivering strong performance in classification problems. They are an important tool in medical diagnostics because of how well they can differentiate between those who are healthy and those who are at danger. Duraisamy et al. [15] achieved an accuracy of 89% in predicting heart disease using Support Vector Machines (SVM), demonstrating the model's effectiveness in analyzing patient data for early diagnosis. In a comparative analysis of various machine learning models for predicting myocardial infarction, Miah, Rahman, and Alam [14] reported that SVM attained an accuracy of 75.01%, highlighting its potential in assessing cardiovascular risk.

SVM's effectiveness in predictive modelling was demonstrated by Rababa, Alshraideh, and Abu-Khalaf, [7], who used them to assess the accuracy of the CDC's heart disease survey data. They were able to achieve up to 80% accuracy while cutting the survey time by 77%. SVM's ability to handle multiclass classification problems in medical diagnostics was highlighted by Wiharto, Kusnanto, and Widayanto [12], who examined the effectiveness of multiclass SVM classification for the diagnosis of coronary heart disorders. Together, these investigations demonstrate the stability and dependability of SVMs in predicting cardiac disease, hence reaffirming their use as useful instruments for clinical judgement.

Some study has directly compared the performance of several models, whereas earlier studies concentrated on the usage of particular models like Random Forest (RF), Decision Trees (DT), and Support Vector Machines (SVM) for heart disease prediction. In their analysis of clinical data from Bangladesh, Hossain et al. [2], discovered that RF performed better than DT and SVM, attaining the greatest accuracy of 90%, indicating its versatility across various demographic groups. Similarly, in a comparison study on myocardial infarction prediction, Miah, Rahman, and Alam [14], found that RF had greater predictive capability, whereas SVM performed moderately (75.01% accuracy). Khan, Anwar, and Sikandar [6] explored a hybrid model combining Random Forest (RF) and Linear algorithms, comparing its performance with Decision Trees and Support Vector Machines. Their results showed that the Random Forest model consistently delivered higher accuracy, demonstrating its reliability for medical diagnostic applications. Hardas, Aush, and Raut [16], examine how data-driven optimization techniques can improve the accuracy and reliability of cardiovascular risk assessment models. Their study highlights the role of advanced analytics, feature selection, and model tuning in enhancing predictive performance and supporting better clinical decision-making. From [17], random forest method of linear equation applied on insurance industry estimates premium using different risk factors from motor data based on policyholders' behavior driving, this is best estimation of pay as you drive.

III. Methodology

Logistic Regression (LR) is a widely used statistical method for handling binary classification tasks. It plays a significant role in medical diagnostics, such as identifying the presence or absence of heart disease, where the outcome variable represents two possible states disease present or absent. In contrast to continuous outputs in linear regression, logistic regression provides a best estimate of the likelihood or probability of some particular outcome or result occurring and thus is more desirable in clinical fields where predicting a disease probability would be important to early detection and treatment. The fundamental idea of logistic regression lies in the use of the logistic, or sigmoid, function, which transforms input variables into probabilities that fall between 0 and 1. The function is expressed as

$$P(Z=1\,|\,Y)=\frac{1}{1+e^{-(\theta_0+\theta_1Y_1+\theta_2Y_2+\ldots+\theta_nY_n)}} \tag{1}$$

Where probability of Z = 1given Y represents the probability of heart disease given the input features Z, β_0 is the intercept term, $\theta_1,\theta_2,\ldots,\theta_n$ are the regression coefficients assigned to each predictor variable, $Y_1,Y_2,\ldots,Y_n$ represent the independent variables such as age, levels of cholesterol and BP. The model is trained by optimizing the log-likelihood function, which maximizes the probability of correctly classifying the observed data.

$$L(\beta)=\sum_{i=1}^{m}[z_i\log(p(Z_i))+(1-z_i)\log(1-p(Z_i))] \tag{2}$$

Where z_i denotes the actual class label (0 or 1), p (Z_i) is the predicted probability, The optimization is performed using techniques such as gradient descent or Newton- Raphson methods. In logistic regression, the sigmoid function is essential because it converts the output of linear regression into a probability value that is limited to a number between 0 and 1. Without the sigmoid function, logistic regression wouldn't be able to output probabilities, which are needed for classification. It's described as

$$sigmoid(z)=\frac{1}{1+e^{-z}} \tag{3}$$

In logistic regression, the variable z represents a linear combination of input features and their associated coefficients. Observations with predicted probabilities equal to or above 0.5 are classified as having heart disease, while those with lower anticipated probabilities are classified as not having it. This threshold is often set at 0.5. Age, gender, resting blood pressure, cholesterol, maximum heart rate, fasting blood sugar, ECG results, exercise-induced angina, ST depression and slope, number of major vessels, and thalassaemia status are among the clinical indicators included in the Heart Disease Statlog dataset from Kaggle, which is used to train the model. The dataset is pre-processes before training in order to encode category data, scale numerical variables, and handle missing values. Stratified sampling is then used to separate it into training (80%) and testing (20%) subsets while maintaining the class distribution. By minimising a cost function, logistic regression finds the best-fit coefficients for the predictor variables during model training. These coefficients indicate how each independent feature influences the probability of heart disease. Once trained, the model assigns a probability to each test sample, which is then translated into a binary outcome using a 0.5 threshold classifying instances as either having or not having heart disease.

$$\text{Prediction} = \begin{cases} 1, & \text{if } p(Y = 1 \mid X\ ) \geq 0.5 \\ 0, & \text{otherwise} \end{cases}$$

This classification rule guarantees that a label identifying the presence or absence of cardiac disease is given to every patient. A crucial prediction method for detecting instances of heart disease is logistic regression. The probabilistic approach together with interpretability and simple implementation structure makes logistic regression an attractive solution for medical diagnostic purposes. Healthcare professionals can use multiple risk factor evaluation through this method to determine outbreak likelihood and launch preventive treatments. Linear relationships exist between predictors and log-odds in logistic regression although this model faces limitations in its assumption. Although logistic regression maintains its status as an effective method for healthcare predictive modelling.

Logistic Regression belongs to the family of Generalized Linear Models (GLMs) and is specifically used when the dependent variable follows a binomial distribution. It is perfect for classification problems since it maps anticipated values to a range between 0 and 1 using the logit link function. By transforming the probability through the logit function, it establishes a linear connection between the independent variables and the transformed response. The binary classification GLM is expressed as:

$$g(E(Y| X)) = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n \tag{4}$$

where, $g(E(Y|X))$ is the link function that transforms the expected value of the response variable, Y is the dependent variable (heart disease status 0 or 1), $X_1, X_2, \ldots, X_n$ are the predictor variables (age, cholesterol levels, blood pressure, etc.), $\beta_0, \beta_1, \beta_2 ..., \beta_n$ are the regression coefficients.

In order to determine a prediction model's dependability and effectiveness in diagnosing heart disease, it is crucial to evaluate its performance. A model that works well in a controlled data set but not in real-world settings can have disastrous implications, particularly in medical decision-making. A model that performs well in a controlled dataset but fails in real-world applications can have serious consequences, especially in medical decision-making. In order to demonstrate how the LR model may be applied practically in healthcare settings, this section examines the performance metrics that are used to evaluate it. To assess the model's performance, several important performance metrics were employed. Each of the measurements provides a different insight into how effectively the model can differentiate between patients with and without heart disease.

- Accuracy is defined as the percentage of accurately predicted instances relative to all instances. Despite being a commonly used evaluation statistic, it might be deceptive when there is a class imbalance and one class is substantially outnumbered by the other. Without successfully recognising the minority (positive) examples, a model may attain high accuracy in such situations by favouring the dominant class.
- The precision measures how many correctly anticipated positive cases there are in comparison to actual positive cases. High precision means fewer false positives in a medical setting, which is especially crucial to prevent needless anxiety, therapy, and medical procedures for patients who do not have cardiac disease. A highly precise model guarantees that if it predicts heart disease, it is quite certain to do so.
- Recall measures the extent to which the model is able to pick up true cases of heart disease. This measure is especially important in medical diagnosis since missing a case of heart disease in a patient who actually has it can have serious health implications. A high-recall model guarantees that the majority of true cases are picked up, even if this means misclassifying some healthy people at times.
- A model's performance can be fairly evaluated with the F1-score, which is determined by taking the harmonic mean of precision and

recall. In medical datasets, which often exhibit an imbalance between positive and negative instances, the F1-score accounts for both false positives and false negatives equally. This helps ensure the model doesn't achieve high performance in one metric (such as precision) at the cost of another (like recall).

- By graphing the true positive rate versus the false positive rate, the AUC-ROC score assesses a model's capacity to differentiate between people with and without heart disease. A higher AUC-ROC value indicates stronger discriminatory power, meaning the model is more effective at identifying high-risk patients while minimizing incorrect classifications. This metric helps balance specificity and sensitivity based on clinical demands, which makes it particularly useful for selecting an appropriate categorisation threshold.
- Through the consideration of these performance metrics, we are able to have a complete view of the strengths and weaknesses of the model. For heart disease prediction, a perfect model would have a fine line to walk properly identifying the risk cases while limiting unnecessary diagnoses. This ensures optimal use of medical resources and that patients get timely and adequate care.uthors and Affiliations

IV. Analysis and interpretation

The model's output is examined in this part along with its applicability to the prediction of heart disease. The analysis of correlations between various risk factors and the occurrence of heart disease reveals hidden patterns and trends in the data. By interpreting these results, one can gain important knowledge about the model's applicability and function in assisting with medical risk assessment. Furthermore, future research on the detection of cardiac disease may benefit from the findings, which could improve prediction accuracy.

Figure 1 heatmap visually illustrates the correlations among various features, shedding light on factors significantly associated with heart disease risk. The range of correlation coefficients is -1 to 1, where a negative number suggests an inverse relationship and a positive value shows a direct association. Age and heart disease have a moderately positive connection, which is consistent with known medical data that the chance of having heart disease rises with age. A strong negative correlation with maximum heart rate indicates that individuals with heart disease often

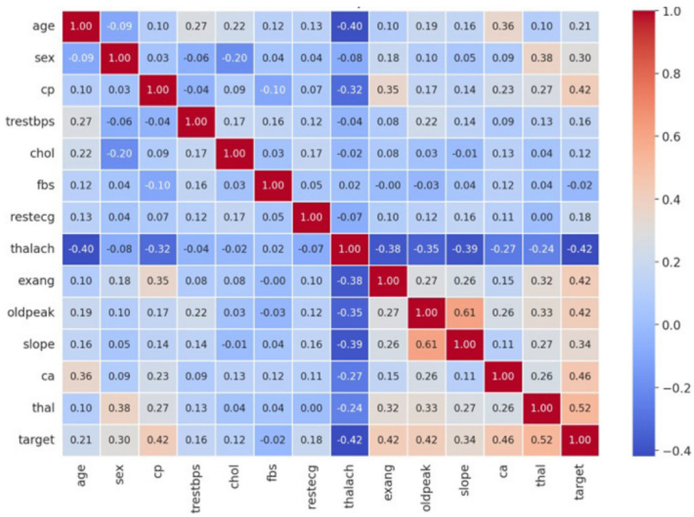

Fig. 1

Correlation Heatmap

exhibit lower peak heart rates during exertion. Additionally, a positive correlation between chest pain type (cp) and heart disease risk shows that more severe chest pain types, such as typical angina, are strong indicators of heart disease. ST depression, a marker of heart stress during physical activity, also demonstrates a significant positive correlation, pointing to impaired cardiac function under strain. Finally, the presence of exercise-

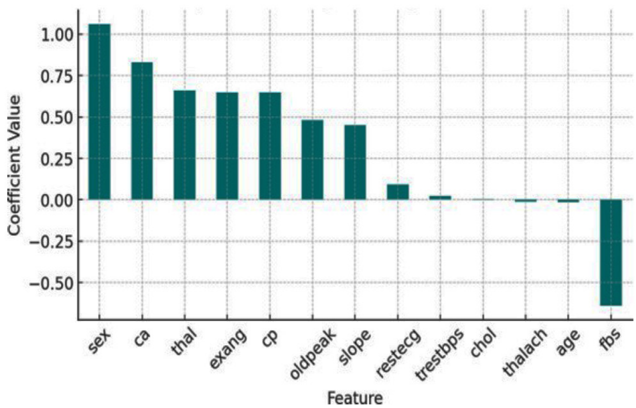

Fig. 2

Logistic Regression Coefficient Values

induced angina is positively linked with heart disease, highlighting the diagnostic value of exercise stress tests.

Features like thalach, cp, oldpeak, and exang are strong indicators of heart disease and should be prioritized in predictive models, as they provide significant insights into a patient's cardiovascular health and response to physical exertion. The inverse relationship of thalach (maximum heart rate achieved) suggests that heart rate performance is crucial for risk assessment, as lower maximum heart rates are often associated with reduced cardiac efficiency and a higher likelihood of heart disease. The positive correlation of oldpeak and exang highlights the impact of exercise stress on heart health, emphasizing how deviations in ST depression (oldpeak) and induced angina (exang) during physical activity can serve as important diagnostic markers for identifying individuals at higher cardiovascular risk.

From the Figure 2, Sex (male), number of major vessels (CA), thalassemia (Thal), exercise-induced angina (Exang), and chest pain type (CP) significantly increase the likelihood of heart disease. Oldpeak (ST depression), slope of ST segment, and age contribute to risk but are less influential than the top predictors. Resting ECG (Restecg) and blood pressure (Trestbps) show minimal effect on heart disease risk. Maximum heart rate achieved (Thalach) reduces heart disease risk, while fasting blood sugar (Fbs) unexpectedly has a negative association. Overall, sex and CA are the strongest predictors, while age plays a role but is not the most significant factor.

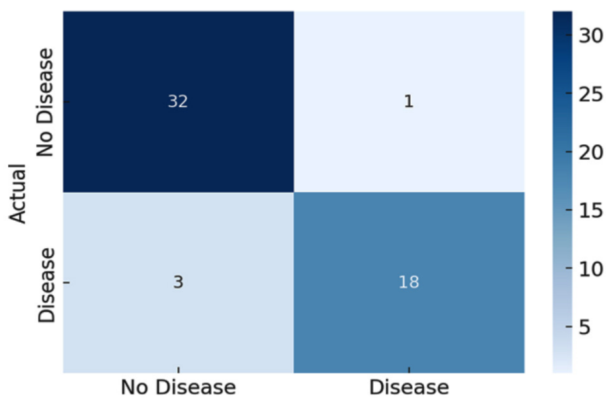

Fig. 3

Confusion Matrix

By describing the proportion of accurate and inaccurate predictions in each class, a confusion matrix provides a thorough assessment of a classification model that goes beyond simple accuracy. As illustrated in Figure 3, the matrix reveals that the model accurately identified 18 individuals with heart disease (True Positives) and correctly classified 32 healthy individuals (True Negatives). There was 1 instance where a healthy person was incorrectly labeled as having heart disease (False Positive), and 3 cases where heart disease was not detected in affected individuals (False Negatives). In medical diagnostics, false negatives are particularly concerning, as missed diagnoses can have serious health implications, while false positives, although less dangerous, may cause undue anxiety and lead to unnecessary testing. The model achieves an accuracy of 92.59%, indicating it correctly predicts the condition in nearly 93% of cases. It also demonstrates a precision of 94.7%, showing that the vast majority of positive predictions are accurate. With a recall rate of 85.7%, the model successfully identifies most true cases of heart disease. The F1-score of 89.9% reflects a strong overall balance between precision and recall.

The ROC curve, which shows the model's high ability to distinguish between people with and without heart disease, has an AUC value of 0.9495 (about 95%) in Figure 4. The model produces very precise and reliable predictions when its AUC is near 1.0. The diagonal reference line on the ROC map indicates a random guess (AUC = 0.5), which is far worse than this performance. The model's high sensitivity, which minimises false negatives while accurately recognising the majority of heart disease cases, is highlighted by the curve's strong early climb. The model's high AUC value indicates a well-balanced trade-off between sensitivity and

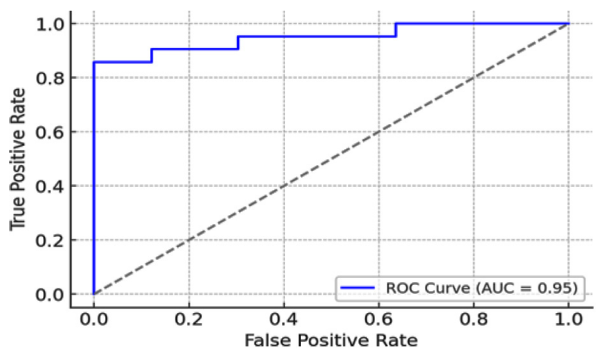

Fig. 4

ROC Curve

specificity, which makes it especially appropriate for clinical settings where prompt and precise diagnosis is crucial.

V. Conclusion and suggestion to stakeholders

In this study, the Heart Disease Statlog dataset was utilised to test a Logistic Regression (LR) model for heart disease prediction. The methodology comprised extensive data exploration via exploratory data analysis (EDA), model performance testing, correlation analysis, and feature importance rating. With a 92.59% accuracy rate and an AUC-ROC score of 0.95, the logistic regression model demonstrated that it was a powerful predictive tool for differentiating between those with and without heart disease. Sex, the number of main vessels (ca), thalassaemia status (thal), exercise-induced angina (exang), chest pain type (cp), and ST segment depression during exercise (oldpeak) were among the key indicators of heart disease that were emphasised in the findings. In contrast, variables such as resting ECG results (restecg), cholesterol levels (chol), and fasting blood sugar (fbs) were found to have lower predictive significance, indicating a lesser impact on heart disease prediction within this particular dataset.

The logistic regression model's excellent accuracy (92.59%), precision (94.7%), and recall (85.7%) demonstrated its great performance and dependability. The AUC-ROC score of 0.95 further supported the model's ability to discriminate between people with and without heart disease across a range of probability thresholds. The interpretability of logistic regression, in contrast to black-box models like deep learning or ensemble approaches, is one of its main advantages. The coefficient-based feature importance analysis shed important light on the relative contributions of each risk factor, showing that fasting blood sugar (fbs) had a negative effect on the likelihood of heart disease, while sex, the number of major vessels (ca), and the results of a thalassaemia test (thal) had the greatest positive effects. Additionally, the model achieved a well-balanced tradeoff between accuracy and recall, capturing the majority of true heart disease cases with a recall of 85.7% and assuring minimal false positives (incorrectly predicting heart illness) with a precision of 94.7%. In medical applications, where inaccurate forecasts may result in misdiagnosis or needless therapies, this balance is essential. Additionally, the project used rigorous statistical validation procedures, such as AUC-ROC interpretation, accuracy assessment, precision-recall evaluation, and confusion matrix analysis. The confusion matrix findings further confirmed the model's accuracy in predicting heart illness, showing a low error rate of 18 true positives (TP), 32 true negatives (TN), 1 false positive (FP), and 3 false negatives (FN).

Future research can leverage real-time data from wearables like smartwatches and fitness trackers to monitor vital signs, improving risk prediction and proactive intervention. Streaming machine learning models could detect anomalies and long-term patterns beyond logistic regression's capabilities. This study's logistic regression model, with an AUC-ROC of 0.95, highlighted key predictors such as sex, number of major vessels (ca), and thalassemia test results, while cholesterol had minimal impact, underscoring the need for data-driven feature selection. However, limitations such as dataset size, biases, and linear model constraints suggest the need for larger, more diverse datasets and advanced feature selection methods like Principal Component Analysis (PCA). Exploring non-linear models like Random Forest, GBM, or deep learning could enhance accuracy by capturing complex interactions. Future studies could also integrate genomic data and AI-driven diagnostics, enabling personalized risk assessment. Actuarial applications may benefit from predictive modeling to refine health risk assessments and dynamic insurance pricing based on lifestyle and real-time health data. Wearable-driven wellness programs could incentivize healthier habits with personalized interventions and insurance discounts. Despite its limitations, this study reinforces logistic regression's role in clinical diagnosis, offering a foundation for actuarial science, clinical decision-making, and preventive healthcare, ultimately improving patient outcomes.

References

[1] R. Torthi, A. D. K. Marapatla, S. Mande, H. K. V. Gadiraju, and C. Kanumuri, "Heart disease prediction using random forest based hybrid optimization algorithms," *Int. J. Intell. Eng. Syst.*, vol. 17, no. 2 (2024).

[2] S. Hossain, M. K. Hasan, M. O. Faruk, N. Aktar, R. Hossain, and K. Hossain, "Machine learning approach for predicting cardiovascular disease in Bangladesh: Evidence from a cross-sectional study in 2023," *BMC Cardiovasc. Disord.*, vol. 24, no. 1, p. 214 (2024).

[3] P. Deepika, "Heart disease prediction using classification with different decision tree techniques," *Int. J. Eng. Res. Technol.,* vol. 12, no. 5, pp. 45–50 (2023).

[4] A. K. Pandey, P. Pandey, K. L. Jaiswal, and A. K. Sen, "A heart disease prediction model using decision tree," *IOSR J. Comput. Eng.*, vol. 25, no. 4, pp. 10–17 (2023).

[5] S. Narayan and J. Gobal, "Optimal decision tree fuzzy rule-based classifier for heart disease prediction using improved cuckoo search algorithm," *Int. J. Intell. Eng. Syst.,* vol. 16, no. 3, pp. 112–121 (2023).

[6] Z. Khan, S. Anwar, and G. Sikandar, "Heart disease prediction using hybrid random forest and linear model," *Int. J. Emerg. Eng. Technol.*, vol. 2, no. 1, pp. 6–12 (2023).

[7] M. Rababa, H. Alshraideh, and J. Abu-Khalaf, "Evaluating the accuracy of the CDC heart disease survey data using support vector machines," *Comput. Biol. Med.*, vol. 153, p. 106512 (2023).

[8] N. M. Lutimath, N. Sharma, and B. K. Byregowda, "Prediction of heart disease using random forest," in *Proc. 2021 Emerging Trends in Industry 4.0 (ETI 4.0)*, pp. 1–4 (May 2021).

[9] L. Yang, H. Wu, X. Jin, P. Zheng, S. Hu, X. Xu, W. Yu, and J. Yan, "Study of cardiovascular disease prediction model based on random forest in eastern China," *Sci. Rep.,* vol. 10, no. 1, p. 5245 (2020).

[10] S. Akila and S. Chandramathi, "A hybrid method for coronary heart disease risk prediction using decision tree and multi-layer perceptron," *Indian J. Sci. Technol.,* vol. 8, no. 34, pp. 1–7 (2015).

[11] V. Sabarinathan and V. Sugumaran, "Diagnosis of heart disease using decision tree," *Int. J. Res. Comput. Appl. Inf. Technol.,* vol. 2, no. 6, pp. 74–79 (2014).

[12] W. Wiharto, H. Kusnanto, and R. Widayanto, "Performance analysis of multiclass support vector machine classification for diagnosing coronary heart diseases," *Health Inform. J.,* vol. 21, no. 3, pp. 199–211 (2015).

[13] E. J. Benjamin, S. M. Al-Khatib, P. Desvigne-Nickens, A. Alonso, L. Djoussé, D. E. Forman, A. S. Go, P. G. Hylek, S. L. Johnson, J. M. Mad-

dox, M. E. McCabe, J. P. Piccini, and E. Z. Soliman, “Research priorities in the secondary prevention of atrial fibrillation: A National Heart, Lung, and Blood Institute virtual workshop report,” *J. Am. Heart Assoc.,* vol. 10, no. 16, p. e021566 (2021), doi: 10.1161/JAHA.121.021.

[14] M. A. Miah, M. M. Rahman, and S. Alam, “Comparative analysis of machine learning models for myocardial infarction prediction,” *J. Biomed. Inform.,* vol. 136, p. 104198 (2023).

[15] B. Duraisamy, R. Sunku, K. Selvaraj, V. V. R. Pilla, and M. Sanikala, “Heart disease prediction using support vector machine,” *Multidiscip. Sci. J.*, vol. 6, p. e2024ss0104 (2024).

[16] B. M. Hardas, M. G. Aush, and V. Raut, “Data-driven optimization strategies for enhanced cardiovascular risk assessment,” *J. Stat. Manag. Syst.* vol. 27, no. 2, pp. 315–325 (2024), doi: 10.47974/JSMS-1257

[17] S. B. Latha and S. Prettha, “Technology adoption alters the insurance industry’s competitive landscape in India,” in *Generative AI Foundations, Developments, and Applications*, IGI Global Scientific Publishing, pp. 219–250 (2025).

Received June, 2025