AOI: 10.100.234513.0255
13th international conference on Microelectronics, Circuits and Systems, Micro2026

Heart Disease Prediction Using Logistic Regression: A Machine Learning Approach

Utkarsh Kesharwani, Jhalak Gour

School of Computer Science and Engineering.
Galgotias University, Greater Noida, India.
Corresponding Author: Utkarsh.22scse1011164@galgotiasuniversity.edu.in

ABSTRACT

Heart disease is the number one cause of death across the globe with millions of deaths annually and is a serious challenge to healthcare systems in the world today. The early and efficient diagnosis is vital in guaranteeing the timeliness of treatment and the management of the disease which may further decrease the overall mortality rate. In this study, we use the tools of Machine Learning (ML) namely, Logistic Regression (LR) to estimate the probability of heart disease through several clinical and physiological variables including age, blood pressure, cholesterol, and heart rate. The main hypothesis of the research is to test the predictive power and clinical interpretability of the Logistic Regression model in order to detect people with the risk of developing heart disease. We used a publicly available heart disease dataset to train and validate the LR model to be able to classify patients as either being at risk or not. The model has an accuracy of 85 which shows that it has a good balance between sensitivity and specificity. According to our results, Logistic Regression can provide a valid, clear, and computationally greedy way of predicting heart disease. Moreover, we ran LR performance against other machine learning models like the Decision Trees, the Random Forests, and the Support Vector Machines. Some of the models had a slight improvement in the accuracy, but LR had a better interpretability- which is critical in clinical decision-making. This study highlights the importance of interpretable ML models in healthcare since they help medical professionals to comprehend risk factors better, assist in diagnostic work, and optimize patient outcomes..

Keywords: Heart disease, machine learning, Logistic Regression, healthcare, predictive modeling, risk assessment, clinical data, machine learning.

I. INTRODUCTION

Heart disease is the number one cause of death in the world according to World Health Organization. According to (WHO), which accounts for 31 percent of the total deaths in the world. With the ever-growing trends of cardiovascular diseases (CVDs), it is important to determine the at-risk individuals to prevent the disease early enough to intervene. Traditional methods of diagnosing CVDs, which are largely based on invasive procedures and medical expert knowledge, are often resource and time-consuming. Consequently, forecasting the heart disease based on non-invasive data and machine learning has become popular over the last several years. The existing studies explore the prediction of CVDs using clinical variables through the application of LR, ML technique which is widely used in binary classification. LR is easy to use, understandable, and time-saving making it an apt model to use in healthcare applications, where clinical decision making relies on knowledge of model outputs.

II. LITERATURE SURVEY

Some of the ML methods used in predicting heart disease include classical statistical predictors. to multilayer deep learning. In the early works in the field of heart disease prediction, traditional study was used. Statistical methods. Indicatively, Detrano et al. (1989) used Logistic Regression on clinical data. to forecast heart disease with a 77% accuracy. Similarly, the Framingham Heart Study constructed by. A clinical risk predicting model of cardiovascular risks that uses clinical variables such as age, blood pressure. Wilson et al. (1998) have described (BP), and cholesterol. With the advancement in science of ML, more complex algorithms started to be studied by researchers. Kora and Kalva (2015) used Naive Bayes, K-Nearest Neighbors (KNN), and Support Vector Machines.

(SVMs) in predicting CVDs. Although SVM was considered as good with an accuracy of 84 the accuracy of the model with black was 84. box nature was problematic in terms of interpretability. A comparison of decision was done by Soni et al. (2011). trees (DT), neural networks and Naive Bayes with the neural networks achieving a perfect accuracy of. 85%. Nevertheless, there is criticism of complex models that they are not as interpretable to facilitate clinical use. Application. Later papers were on ensemble techniques and deep learning. Gupta and Verma According to (2018), a hybrid method of Random Forests and Genetic Algorithms was introduced and yielded better results. An accuracy of 87%. Convolutional Neural Networks (CNNs), were used by Rajasekaran et al. (2020). With a 91% accuracy, however, they are computationally costly and are hard to interpret. Although more complex models have advantages, the interpretability of the Logistic Regression has. Remained its main strength. Tewari et al. (2018) discussed the practicality of the Logistic Regression. in the prediction of cardiovascular events, it is important to mention that its simplicity can be seen clinically in an unequivocal way. Thus, although

ISBN: 978-81-689089-0-1

improved accuracy can be provided by slightly better machine learning models, Logistic Regression is still a useful and clear weapon of predicting heart diseases, especially when there is involvement of clinical decision-making.

III. METHODOLOGY

A. ***Dataset:*** The up-to-date studies have used Heart Disease Dataset on the UCI ML Repository. Dataset comprises 303patient records, which involve 14clinical parameters that are used to predict the presence of CVDs. Target variable is binary (1=heart disease, 0=no heart disease).

B. Features:

The features used in CVDs prediction are:

- Age: Patient's Age.
- Fasting Blood Sugar (fbs): Refers to fbs>120mg/dl.
- Major Vessels (ca): The number of major vessels with a fluoroscopic color.
- Max Heart Rate (thalach): Farthest heart rate.
- Old peak: ST depression as a result of exercise.
- Resting Electrocardiographic Results (restecg): Resting ECG Results.
- Serum Cholesterol (chol): Cholesterol.
- Sex: Patient's Gender.
- Chest Pain Type (cp): types of chest pain experienced In this condition, a person experiences chest pain during physical activity.
 <|human|>Exercise-Induced Angina (exang): In case of angina exercises 100mmHg at rest 100mmHg at rest.
- Slope: exercise ST segment slope of peak. Thalassemia (thal): Blood condition demonstrative of heart disease risk.

C. Data Preprocessing Before training model we did preprocess in the following way. steps:

- Missing values: Missing values have been replaced by the median value where necessary.
- Normalization: Cholesterol and BP measures were normalized in order to normalize data.
- Train-Test Split: To do the assessment, dataset is split into 80% training and 20% testing.

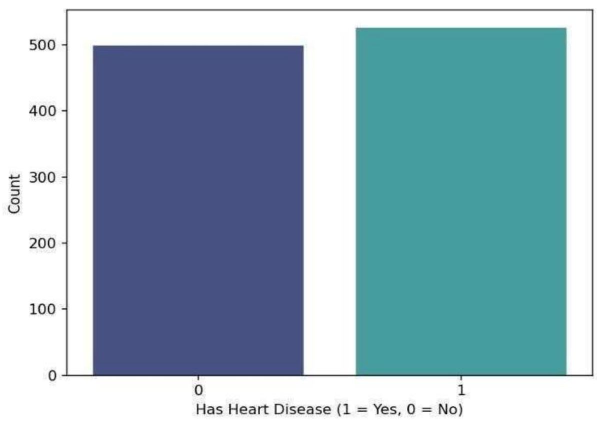

Fig. 1. Distribution of Heart Disease in Dataset

D. Logistic Regression Model:

In LR binary classification algorithm, probability is dependent on the input variables. Linear the input information is transformed into a probability between 0 and 1 by combinations of input information through logistic. Function. LR was used to establish whether or not the patient was at risk of CVDs in the present study.

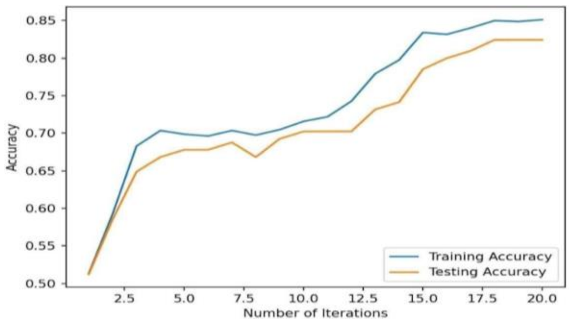

Fig.2. Training Progress of Logistic Regression Model

IV. RESULTS AND DISCUSSION

A. Model Performance:

LR model achieved following metrics on the test set:

TABLE I
PERFORMANCE METRICS OF THE PROPOSED MODEL

Metric	Value
Accuracy	85%
Precision	83%
Recall	87%
F1-Score	85%

Based on these findings, LR model is a good tool in forecasting CVDs. Model accurately differentiates in patients that possess heart disease and those that do not with an accuracy of 85 percent. Also, with a balance between recall and precision, the model is capable of reducing false. Positives and also performing well in identifying the patients at risk.

B. Advantages of the Logistic Regression. Interpretability:

- The LR coefficients provide understandable data on ways in which characteristics. Including age and cholesterol affect the risk of CVDs.

- Probability Estimation: Model gives probabilities on every prediction, which enable clinicians. To assess the level of risk.

- Computational Efficiency: LR may be used in real time since it is computing efficient. clinical prediction

C. Limitations:

In spite of its strong points, LR has weaknesses:

ISBN: 978-81-689089-0-1

- Linear Boundaries: Linear model supposes linear relationships in the features and targets, perhaps. Overlooking complicated interrelationships.

- Sensitivity to Data Outliers: LR may be influenced by large data values and may be skewed. Predictions..

V. COMPARATIVE ANALYSIS

Although LR is appropriate in predicting cardiac disease, more complicated models such as Random Forests can be used. and Neural Networks often achieve greater levels of accuracy. Nevertheless, these models usually need more computing power and cannot be interpreted as easily, so they cannot be practically used in. healthcare settings. Indicatively, although the researchers attained 91% accuracy in a study conducted by Rajasekaran et al. (2020). The complexity of the CNN makes it hard to comprehend the way that predictions are obtained. Conversely, the ensemble method used by Gupta and Verma (2018) was 87% accurate, yet it was less precise. the simplicity and interpretability which render the Logistic Regression appropriate in the real-life clinical situation. Therefore, in spite of the prospects of studying more complicated models, LR offers. Reasonable approach to methodology combining sufficient accuracy and interpretability accuracy.

VI. CONCLUSION AND FUTURE WORK

As this study has shown, LR can be used as a valuable clinical prediction tool of CVD. It is a strong, easy-to-interpret model that strikes a balance between the accuracy and simplification and is thus suitable. in the domain of healthcare. Although more complicated models are capable of a higher accuracy, LR is still. a great option in cases where transparency and clinical understandings are imperative. The next stage of work may be done to improve the models using more features or fusing. Other techniques in combination with LR. Also, the shortcoming of linearity in LR. could also be improved to increase its predictive accuracy.

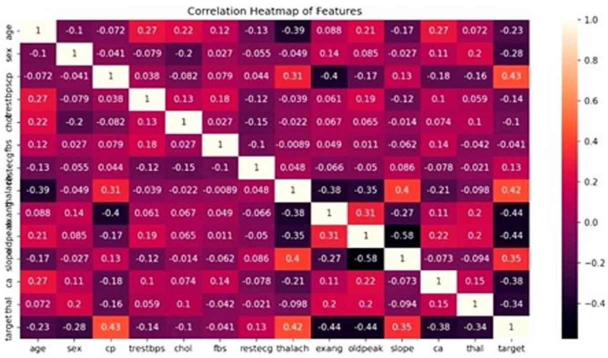

Fig. 3. Correlation Heatmap of Clinical Features Used for Cardiovascular Disease Prediction..

REFERENCES

[1] Detrano, R., et al., "Heart Disease Data Set," UCI Machine Learning Repository, 1988.

[2] Wilson, P.W.F., et al., "Prediction of Coronary Heart Disease Using Risk Factor Categories," Circulation, 1998.

[3] Kora, P., & Kalva, S.K., "Improved SVM Based Heart Disease Prediction System," Journal of Theoretical and Applied Information Technology, 2015.

[4] Soni, J., Ansari, U., Sharma, D., & Soni, S., "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction," International Journal of Computer Applications, 2011.

[5] Gupta, S., & Verma, A.K., "Heart Disease Prediction using Hybrid Random Forest with a Genetic Algorithm," International Journal of Computer Science and Information Security, 2018.

[6] Rajasekaran, M., et al., "Heart Disease Prediction Using CNN," Journal of Computational and Theoretical Nanoscience, 2020.

[7] Tewari, S., et al., "Logistic Regression for Cardiovascular Risk Prediction: A Tool for Early Diagnosis," Journal of Medical Systems, 2018.

Cite this article as:

Utkarsh Kesharwani, and et. al. "Heart Disease Prediction Using Logistic Regression: A Machine Learning Approach", Proceedings of 13th international conference on Microelectronics, Circuits and Systems, Micro2026 (Vol-II)

Displayed as online on 17th July 2026.

Link: http://actsoft.org/science/act2026-pro/496-micro2026.pdf

ISBN: 978-81-689089-0-1