RCHUB: JOURNAL OF COMPUTATIONAL INTELLIGENCE SCIENCE AND ENGINEERING (JCISE)
VOL: 2, ISSUE 4, OCTOBER - DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/001

Early Diagnosis of Cardiovascular Disease Using Logistic Regression

P.Mangaiyarkkarasi[1] and S.Saraswathi[*]

1. Research Scholar, Nehru Arts and Science College(A), Coimbatore – 641105 Tamilnadu , India.

Author Email Address: mangaisivaphd@gmail.com

2. *Dean Academic affairs, Nehru Arts and Science College(A), Coimbatore – 641105 Tamilnadu, India.

Corresponding Author Email Address: nascsaraswathi@nehrucolleges.com

Abstract

Cardiovascular disease is one of the most life-threatening diseases and accounts for nearly 17 million deaths worldwide. Early diagnosis plays a crucial role in providing timely treatment and reducing mortality. Several machine learning and deep learning techniques are available to classify the presence or absence of cardiovascular disease. In this study, a Logistic Regression (LR) model is applied to the UCI dataset for heart disease classification. To enhance the model performance, data preprocessing steps such as dataset cleaning, missing value handling, and feature selection were carried out. Feature selection was performed based on the correlation between each feature and the target variable, and highly positively correlated features were selected. The dataset was then divided into training and testing sets using different split ratios: 90:10, 80:20, 70:30, 50:50, and 40:60. Among these, the 90:10 split ratio achieved the best performance. The proposed Logistic Regression model attained an accuracy of **87.10%**.

Keywords: Cardiovascular disease, UCI data set, Future selection, Logistic Regression

1. Introduction

Today the greatest challenge to medical industry to provide higher level facility to health infrastructure to diagnose the disease in the ini- tial day and give timely treatment to improve the quality of life through quality of service. Around 31% of mortality occurs world due to cardiac disease [1 , 2]. The developing and under developing countries lacks in infrastructure and

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/001

technologies, infrastructure and doctors to predict the disease in early stage to avoid complications reduce mortality. The growth of Information and telecommunication technology has benefited from rich to poor patients by providing real time information to the pa- tients with lower cost of diagnose and monitoring the patients' health. This has increase in detail health records of the patients dramatically. The vast medical records are available to the research. The medical in- dustry faces enormous challenges in using the huge medical data. The vast amount of data is transformed to obtain valuable and accurate in- formation speedily by machine. Thus, machine learning is the important area. The highly useful machine learning models used to discover the hidden pattern and correlation among features in the dataset [3, 4]. The medical dataset is inconsistent and redundant, appropriate preprocess- ing is pivot step [5]. Various researcher has included risk of different feature the most prevalent are 14 features. Since the feature selection become an important part of the study, based on the feature selection the model increases or decrease the prediction accuracy [6]. The cardiac disease can be predicted with the help of machine learning with greater accuracy will help healthcare to diagnose and treat patient in early stage supporting many patients to diagnose disease in short period of time. Thus, saving millions of lives. In this study, Logistic Regression classifier model are applied .Results are compared with existing studies.

2. Problem statement

Presently, the major challenge of the medical industry is to predict the cardio vascular disease with less expensive and more reliable method to avoid the compounding effect of the disease in low income or developing countries. The early detection not only reduce the cost but also improves the quality of life.

3. Research aim and scope of the research

The aim of this research is to develop an efficient way to predict the presence of the cardiovascular disease. The steps as mentioned below.

- The UCI data set is used to predict the disease.
- The Features are selected based on high positive correlation values with the target and used random order of data (without sorting).
- The performance of the mode is evaluated by Five different training and testing ratio of dataset.

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/001

- To check the behavior of the model with low to high training and testing data.

Fig.1. Flow chart of logistic regression model.

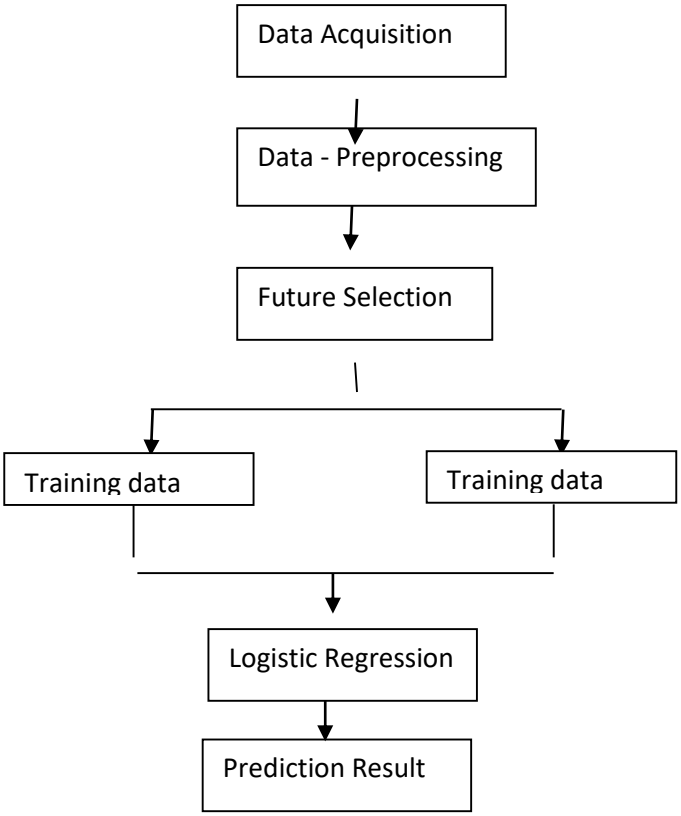


Table1: Features and their correlation with target variable

Feature	**Correlation Value**
Exang	0.436757
Cp	0.433798
Oldpeak	0.430696
Thalach	0.421741
Ca	0.391724
Slope	0.344029

Table 2: Split percentage of training and testing sets

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/001

S. No.	Training Set (%)	Testing Set (%)
1	50	50
2	60	40
3	70	30
4	80	20
5	90	10

Table 3: Accuracy Results of Logistic Regression Classifier

Training : Testing Ratio	Accuracy (%)
90 : 10	87.10
80 : 20	85.25
70 : 30	83.52
60 : 40	81.97
50 : 50	81.58

Table 4: Classification report of logistic regression classifier

Actual / Predicted	Positive	Negative	Precision	Recall	F1-Score	Accuracy (%)
Positive	15	2	0.857	0.857	0.857	87.10
Negative	2	12	—	—	—	—

4. Related work

Firda Anindita Latifah et.al., proposed comparative study of machine learning model namely,logistical regression and random forest for classification of heart disease.

The research done on Framingham data set with 3656 records and training to testing ratio of 70:30. The accuracy of 85.04% was achieved by the model [7].

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/001

Zameer Khan et.al., proposed empirical study of several ML algorithms namely, logistic regression, KNN Classifier, RF, SVM, DecisionTree, Gauss Naïve Bayesian used for classification of cardio vascular disease. The research done on UCI Cleveland dataset. The logistic regres-sion achieved accuracy of 85.71 [8].

Thanuja Nisha diASetal. proposed logistic regression model for classification of heart disease on Framingham dataset with 4238rrecords. The logistic regression achieved the accuracy of 86.66% [9].

Montu Saw et al., proposed logistic regression model to classify the cardiac disease. The research uses Framingham datasets and logistics regression achieved accuracy of 87.02% [10]. Detrano et.al., proposed logistic regression model and obtained 77.0% accuracy [11].

Saba Bashir et al., proposed several ML algorithms namely, logistic regression, decision tree, logistic regression support vector machine etc. The research uses UCI dataset and the logistic regression achieved accuracy of 82.56% and logistic regression support vector machine achieved accuracy of 84.85% [12].

Kannan,Retal.,proposedvariousMLalgorithmsnamely,LR,SVM,RFStochasticGradientBoostingusingUCIClevelanddataset.The logistic regression outperformed other models with best accuracy of 86.51%[13].

Ganesan M et al., proposed four different ML algorithm, SVM, J48, Logistic Regression, Multilayer perceptron using UCI dataset to clas- sify cardiac disease and obtained best accuracy. The logistic regression achieved an accuracy of 83.70% [14] .

Dinesh Kumar G et al., proposed five machine learning algorithm using UCI dataset to classify the cardiac disease. The Logistic Regression achieved overall accuracy of 86.51% [15] .

5. Proposed work

In the proposed system, the analysis of the cardiac disease UCI dataset is carried out using suitable data acquisition, preprocessing by cleaning the data, then using selects all the features

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/001

which have high correlation with the target function. Then logistics regression model was trained and tested for predicting the cardiac disease is present or not. The Fig. 1 shows the workflow to build logistic regression cardiac disease classification model.

The machine learning model was implemented in python 3.7 envi- ronment and used ML libraries sklearn, pandas and matplotlib. Jupyter Notebook to run the code.

5.1. Data acquisition

The cardiac disease dataset obtained from the UCI ML repository. It contains 13 features and 303 records.

Table 5: Comparative results of logistic regression classifier using UCI dataset

Year	Author / Ref.	Tool / Technique	Dataset	Accuracy (%)
2019	[8]	RapidMiner (Logistic Regression)	UCI	82.56
2019	[8]	RapidMiner (Logistic Regression + SVM)	UCI	84.85
2020	[1]	Python (Logistic Regression)	UCI	85.04
2020	[2]	Python (Logistic Regression)	UCI	85.71
2021	This Paper	Python (Scikit-learn, 90:10 Split)	UCI	87.10

5.2. Data pre-processing

Cardiovascular disease UCI dataset is first loaded and then data cleaning and finding missing values was performed on all records. The dataset contains complete information. The attributes of the dataset are multiclass variable in characteristics with double classification.

5.3. Feature selection

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/001

The patient record is identified uniquely by two features of the dataset by sex and age from 13 attributes of the dataset and assign in-dividual ids. The rest of the features consists of medical information. The medical information are vital attributes predicting heart disease. The correlation performed on all 13 attributes with the target value to select the features with high and positive correlation feature as shown in Table 1.

5.4. Splitting dataset

The Table 2 below shows the splitting of the dataset in the following ratios of training and testing set in percentile

5.5 Classification

One of the Simplest and best ML classification algorithm is Logistic Regression. The LR is the supervised ML binary classification algorithm widely used in most application. It works on categorical dependent variable the result can be discrete or binary categorical variable 0 or 1. The sigmoid function is used as a cost function. Sigmoid function maps a predicted real value to a probabilistic value between '0' and '1'.

Logistic Sigmoid function:

$$(x)=1/(1+e\wedge(-x)) \qquad (1)$$

Where, P(x)is probability estimation function a value between 0 and 1,x is input to the probability function (algorithm's prediction value),the mathematical constant is Euler's number and its value is approximately equal to 2.71828 as shown in Eq. (1).

To predict the cardiac disease logistic regression ML model is used, firstly the LR model are trained with five splitting condition and tested with test data for prediction to get the best accuracy and to find themodelsbehavior.Thealgorithmresultscategoryof1and0forpresenceand absences of cardiac disease.

TheLogisticsRegressionModelisdescribedinPseudocode1isusedin both training and testing the data instance.

Input: Feature-selected data

Output: Class label

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/001

1. Read dataset

2. Initialize Logistic Regression model

3. Calculate probability P(1|d)

4. If P(1|d) > 0.5

Class = 1

Else

Class = 0

5. Return classification result

6. Results and discussion

The logistical regression is tested with UCI dataset with five different ratios and their accuracy as shown in the table below. The accuracy of 87.10% obtained by logistical regression for split ratio of training and testing is 90:10. The increasing accuracy of the model by increasing the training is shown in Fig. 2 and accuracy of result in Table 3.

The Logistics Regression increase its accuracy with increasing training by 50% to 90% and 90% training and 10% testing provides highest accuracy of 87.10%.

The Table 4 shows classification report, precision, recall, f1-score and accuracy of LR classifier for UCI dataset with 90% training and 10% testing. The model has precision of 0.857, recall 0.857, F1-score 0.857 and accuracy of 87.10%.

The Table 5 represents the various previous research work carried on Logistic Regression using rapid minor and python on UCI Dataset from the year 2019 to 2021 with accuracy of prediction. 7.

Conclusion

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/001

One of the important areas in industry of medical is prediction of cardiovascular disease, with the available data of the patient to predict the absence and presence of cardia disease. There are several techniques and methods are present for prediction of cardiovascular disease. In this research, Logistic Regression supervised ML algorithm are used to clas- sify the heart disease. To improve the performance, pre-processing of corpus like Cleaning, finding the missing values are done. The vital part is feature selection, which increase the accuracy of algorithm and even focus on the behavior of the algorithm. As the behavior of Logistic regression is as training increases the accuracy of prediction also in- creased. The LR classifier achieved 87.10% of accuracy with training 90% and testing 10%. The results outperformed compared to previous research work. The limitation is only UCI dataset is used in the study and future work try to implement on multiple datasets.

References

[1.] World Health Organization and J. Dostupno, cardiovascular dis- eases: key facts, vol. 13, no. 2016, p. 6, 2016. [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases- (cvds) .

[2.] K. Uyar, A. Ilhan, Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks, Proced. Comput. Sci. 120 (2017) 588–593 .

[3.] N. Kausar, S. Palaniappan, B.B. Samir, A. Abdullah, N. Dey, Systematic analysis of applied data mining based optimization algorithms in clinical attribute extraction and classification for diagnosis of cardiac patients, in: in Applications of Intelligent Optimization in Biology and Medicine, Cham, Switzerland: Springer, 2016, pp. 217–231 .

[4.] M. Shouman, T. Turner, R. Stocker, Integrating clustering with different data mining techniques in the diagnosis of heart disease, J. Comput. Sci. Eng. 20 (1) (2013) 1–10 .

[5.] M.S. Amin, Y.K. Chiam, K.D. Varathan, Identification of significant features and data mining techniques in predicting heart disease, Telemat. Inf. 36 (2019) 82–93 Mar. .

[6.] D. Singh, J.S. Samagh, A comprehensive review of heart disease prediction using machine learning, J. Crit. Rev. 7 (12) (2020) 2020.

[7.] F.A. Latifah, I. Slamet, Comparison of heart disease classification with logistic regression algorithm and random forest algorithm, in: Proceedings of the AIP Conference, 2020, p. 2296, doi: 10.1063/5.0030579 .

[8.] Z. Khan, D.K. Mishra, V. Sharma, A. Sharma, Empirical study of various classification

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/001

techniques for heart disease prediction, in: Proceedings of the IEEE 5th International Conference on Computing Communication and Automation (ICCCA), 2020, pp. 57 62, doi: 10.1109/ICCCA49541.2020.9250852 .

[9.] Nishadi, A.S.T. (n.d.). International journal of advanced research and publications predicting heart diseases in logistic regression of machine learning algorithms by python jupyterlab. https://www.kaggle.com

[10.] [M. Saw, T. Saxena, S. Kaithwas, R. Yadav and N. Lal, Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning, in: saw (Ed.), 2020 International Conference on Computer Communication and Informatics (ICCCI), IEEE, Coimbatore, India, 2020, pp. 1–6. January 22-24 .

[11.] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K.H., Lee, S., & Froelicher, V. (n.d.). International application of a new prob- ability algorithm for the diagnosis of coronary artery disease.

[12.] S. Bashir, Z.S. Khan, F. Hassan Khan, A. Anjum, K. Bashir, Improving heart disease prediction using feature selection approaches, in: Proceedings of the 16th Interna- tional Bhurban3 Conference on Applied Sciences and Technology (IBCAST), 2019, pp. 619–623, doi: 10.1109/IBCAST.2019.8667106 .

[13.] R. Kannan, V. Vasanthi, Machine learning algorithms with ROC curve for predict- ing and diagnosing the heart disease, in: Springer Briefs in Applied Sciences and Technology, Springer Verlag, 2019, pp. 63–72, doi: 10.1007/978-981-13-0059-2_8 .

[14.] M. Ganesan, N. Sivakumar, IoT based heart disease prediction and diagnosis model for healthcare using machine learning models, in: Proceedings of the IEEE (ICSCAN), 2019, pp. 1-5, doi: 10.1109/ICSCAN.2019.8878850 .

[15.] K.G. Dinesh, K. Arumugaraj, K.D. Santhosh, V. Mareeswari, Prediction of cardiovas- cular disease using machine learning algorithms, in: Proceedings of the International Conference on Current Trends towards Converging Technologies (ICCTCT), 2018, pp. 1–7, doi: 10.1109/ICCTCT.2018.8550857 .