Optimizing Logistic Regression Model for Heart Disease Prediction with Machine Learning

2024 Seventh International Conference on Vocational Education and Electrical Engineering (ICVEE) | 979-8-3315-0510-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICVEE63912.2024.10823688

MD. Shameem Ahammed
Electrical and Electronic Engineering
Jatiya Kabi Kazi Nazrul Islam University
Trishal, Mymensingh 2224, Bangladesh
shameem_19102913@jkkniu.edu.bd

Md. Mahbubur Rahman
Electrical and Electronic Engineering
Jatiya Kabi Kazi Nazrul Islam University
Trishal, Mymensingh 2224, Bangladesh
mahbubur@jkkniu.edu.bd

Md. Borhan Uddin
Electrical and Electronic Engineering
Jatiya Kabi Kazi Nazrul Islam University
Trishal, Mymensingh 2224, Bangladesh
borhan_19102936@jkkniu.edu.bd

Zinia Jahan
Electrical and Electronic Engineering
Jatiya Kabi Kazi Nazrul Islam University
Trishal, Mymensingh 2224, Bangladesh
ziniajahan2941@gmail.com

Md. Shake Farid Uddin
Assistant Manager(Software Engineer)
Nuclear Power Plant Company
Bangladesh Limited.
Dhaka, Bangladesh
sfuddin.iit@gmail.com

Sabnam Mushtari
Department of Microbiology
Mymensingh Medical College
Mymensingh, Bangladesh
Sabnam.rumpa1@gmail.com

***Abstract*— Nowadays, heart disease is a serious medical issue, and diagnosis is crucial to providing the proper treatment to the patient. Hence, machine learning (ML)-based early diagnosis of heart disease can reduce time and misdiagnosis. In this paper, we examine the various machine learning models for detecting heart diseases early and propose a method of *optimized logistic regression (OLR)* to make predictions more accurate. The proposed model uses GridSearchCV and five-fold cross-validation to determine the best parameter for the logistic regression (LR) classifier and tune the hyperparameter. For evaluations, we conducted extensive experiments and compared the proposed method with different machine learning methods including *Logistic Regression (LR)*, *Support Vector Machine (SVM), Random Forest (RF), AdaBoost (AB), and Naive Bayes (NB).* The results confirmed that the proposed method improved overall performance with an accuracy of 90.74% and *area under the curve (AUC)* of 0.95.**

Keywords*— heart disease prediction, machine learning, optimized logistic regression, support vector machine, random forest, adaboost, and naive bayes.

I. INTRODUCTION

Cardiovascular diseases (CVDs) are among the most critical health concerns in today's healthcare sector. According to the *World Health Organization (WHO)*, CVDs are projected to cause 23.6 million deaths by 2030, presenting a significant threat to global health. The World Heart Report 2023 highlights key risk factors contributing to the rising CVD death rate, including alcohol consumption, high sodium intake, obesity, smoking, high blood pressure, and physical inactivity. Alarmingly, 80% of CVD-related deaths occur in low- and middle-income countries, exposing stark health inequities in these regions. Many of these countries lack access to essential diagnostic tools, life-saving treatments, and effective healthcare systems to combat CVDs[1][2]. Early detection and prediction of heart disease are crucial in reducing misdiagnosis and improving treatment outcomes. In turn, this can help minimize healthcare costs and complications, enabling individuals to lead healthier, more fulfilling lives.

Several ML techniques have been introduced to predict heart diseases at early stages, providing a promising solution for improved diagnostics. In their study, Bhatt et al. [3] analyzed the accuracy of various ML models in predicting heart disease. To further enhance model performance and convergence, they utilized k-modes clustering for preprocessing and scaling the dataset and found the Multilayer perceptron (MLP) model achieved the highest accuracy (87.28%), outperforming other techniques. However, their study faced challenges with lower accuracy due to the lack of diverse datasets, limiting the generalizability of their results. Similarly, Goutam et al. [4] implemented an ML model for heart disease detection, achieving better accuracy (88.5%) using the RF algorithm. Their study compared RF with other algorithms, including LR, NB, DT, KNN, SVM, and XGBoost, demonstrating the superiority of Random Forest in this context. Karthick et al.[5] developed an early prediction heart disease model using SVM, LightGBM, GaussianNB, XGBoost, LR, and RF classifiers. The RF model achieved highest accuracy of 88.5% for the recursive feature elimination method. There is a need for more diverse datasets to improve the generalizability of these models across different populations and demographics. Nirmala et al.[6] focus on the prediction of heart disease by applying LR, DT, RF, GB, AB, SVM and achieved 88% using the recursive factor removal system via RF method. Handling imbalanced dataset and massive volume of data are the limitation. Mihir et al. [7] explore SVM, GB, DT, RF, and LR classifiers to identify heart disease accurately. Girish et al. [8] discuss heart disease prediction via several deep learning and machine learning techniques. Haoyo et al. [9] published a highly effective prototype for heart disease early identification uses the LR algorithm and gained highest accuracy of 85%. Vatsala et al. [10] compared ML methods for identifying heart disease via LR, DT, NB, KNN, and XGBoost and achieved 86% of accuracy via LR method. Soumyalatha et al. [11] utilized two ML model via LR and RF methods to identify heart disease. The accuracy of the LR model was 85.25%, whereas the RF model had the highest accuracy of 90.16%. Geetha et al. [12] use a LR classifier to estimate the risk of diagnosing heart disease has an 87% accuracy, which is not effectively identify all positive cases and not applicable to broader populations due to the specific dataset used in the analysis. M.Anshori et al. [13] aim to detect heart disease via LR classifiers uses iterative form. At iteration 14, achieved maximum accuracy of 81.3495%. Pratham et al. [14] focus on predicting heart disease using several ML methods, i.e. LR, RF, KNN, DT,

SVM, GBoost, bagging, NV, and AB. The AdaBoost model utilized the highest test accuracy which is 90.20%.

In this paper, we aim to develop a model with improved accuracy by optimizing the features of Logistic Regression to predict heart disease. With this reliable and accurate prediction model, high-risk features can be identified more quickly, allowing for more efficient implementation of prevention methods. We have investigated various ML models in order to identify heart disease earlier and proposed our method of *Optimized Logistic Regression (OLR)* to enhance accuracy in CVD detection. This method has been optimized using the best parameters and hyperparameter tuning via GridSearchCV with five-fold cross-validation.

For evaluation, we conducted extensive experiments and compared the proposed method with others. The *optimized logistic regression (OLR)* model is trained with the best parameter for a dataset from Kaggle, where the train-test split 80:20 and achieved an accuracy of 90.74%. The other algorithm achieved accuracy as follows: logistic regression (LR): 83.33%, Support vector Machine (SVM): 87.03%, Random Forest (RF): 83.33%, AdaBoost (AB): 88.88%, Naive Bayes (NB): 90.74%. The area under the curve (AUC) values gained for optimized Logistic Regression: 0.95, built-in Logistic Regression: 0.90, Support Vector Machine: 0.93, Random Forest: 0.90, AdaBoost: 0.95, Naïve Bayes: 0.92.

In medical applications, the comparison of these model provides valuable insights into how well they perform. In clinical settings, simple models, such as LR model are preferable, because it provides greater transparency, ease of interpretation, explainable decision and when predict is more critical. For real-world applications, need to balance accuracy and interpretability of the model.

The rest of the paper are organized as follows: Section II provides a detailed explanation of the proposed model, including data collection, data pre-processing, exploratory data analysis (EDA), dataset splitting, finding the best parameter, and hyperparameter tuning. Section III implements the SVM(Linear), RF, AdaBoost, and NB classifiers and compares them with the proposed model. Section IV discusses the comparative results. Finally, Section V conclude the paper with some future works.

II. Proposal for Optimized Logistic Regression Method

In this section, the optimized logistic regression method is presented for predicting the heart disease. It is designed by collecting datasets and using them to train our model to determine whether a patient has heart disease or not more accurately. The Fig 1 shows the working procedure of the proposed method.

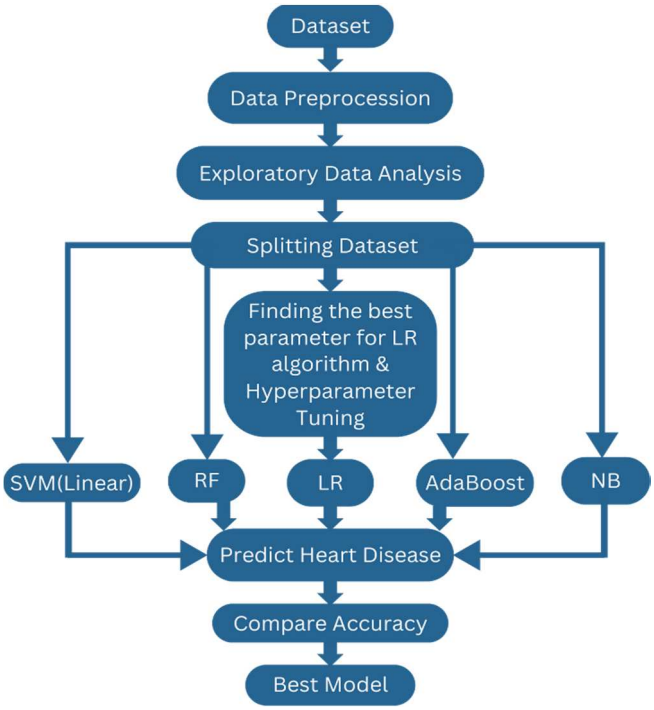

Fig 1: Working procedure of the proposed method

A. Dataset

The dataset is collected from the UCI Machine Learning Repository[15]. It has 270 instances, including 13 features and 1 targeted value. There are two types of targeted value. One is 0, which means a healthy heart, and another is 1 means an infected heart. It contains information about individuals who have and haven't experienced heart disease-causing. The features are age, sex, chest pain(cp), resting blood pressure(trestbps), serum cholesterol(chol), fasting blood sugar(fbs), resting electrocardiogram(restecg) results, maximum heart rate achieved(thalach), ST depression induced by exercise relative to rest(exang), the slop of the peak exercise ST segment(oldpeak), Number of major vessels(ca), thalassemia(thal), and target which means persons have heart disease or not. If data scarcity or imbalance exist, to improve the model performance, data augmentation, transfer learning, cost-sensitive techniques can be employed, whose are ensure more accurate prediction in real-world applications.

Data pre-processing and analysis are commonly used in Python libraries. We imported Pandas (To analyze data), Numpy (used for working with array), matplotlib (serves static, interactive visualization), and Seaborn (to visualize random distributions) libraries.

B. Data Preprocessing

Here, we checked whether our dataset has missing values, duplicate values, or noisy data and found no missing values or duplicate data.

C. Exploratory Data Analysis

In this section, we have seen that our dataset has 270 data where 150 have healthy hearts and the remaining are suffering from heart disease. To understand the targeted values clearly Fig *2* shows the correlation heatmap of each feature with the others.

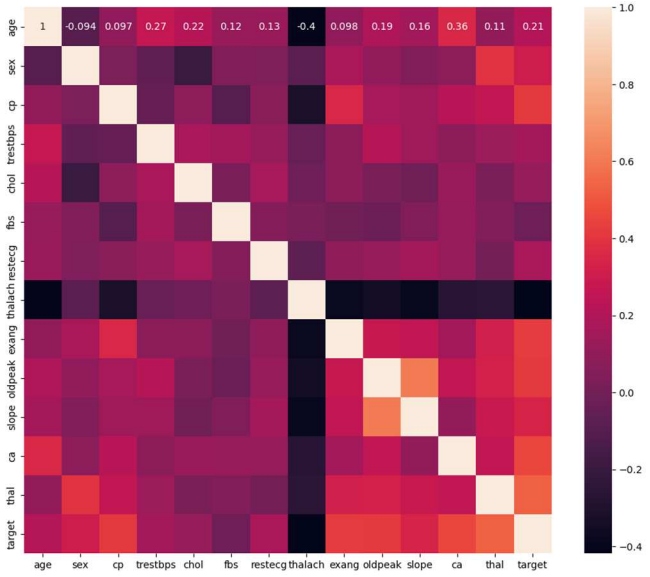

Fig 2: Correlation Heatmap

To visualize the relationship of every feature with each other in the dataset the pair plot is generated in Fig *3*. It combines scatter plots and histograms at a time and gives an outline of correlations and distributions of the datasets. Here diagonal plots are known as histograms which show the distribution of a single feature. The Off-diagonal plots known as scatter plots show the relationship between two features[16].

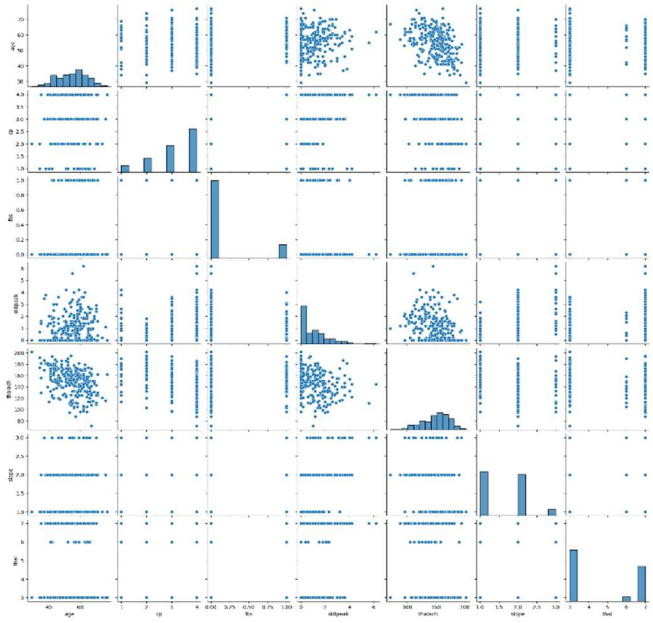

Fig 3: Pair Plot

D. *Splitting Dataset*

In this step, we split our dataset into two parts, i.e. training and testing data. The ratio of train and test data is 80:20. So. to train our model we use 216 data and for testing the sharpness of our system we use 54 data.

E. *Logistic Regression Model*

Logistic Regression is a supervised machine learning algorithm that classifies probability values between 0 and 1. The logistic function "S" shaped assumes two maximum values 0 or 1. It is also known as the mathematical Sigmoid function used to map probabilities. Here, Equation 1 shows the sigmoid function where z denoted as input of the function.

$$Y(z) = \frac{1}{1+e^{-z}} \ldots\ldots\ldots\ldots\ldots\ldots 1$$

and z = wX + b

'X' is the input feature, uses all values from the dataset expect target values. 'w' is the weights which decide how much influence the input will have on the output and 'b' is the bias that defines the difference between the average prediction and correct value which we are trying to predict [17]. The algorithm is simple, interpretable, make linear relationship within the data and handle limited non-linearities.

1) Choose the optimal parameter: The necessity of finding the best parameter for the logistic regression algorithm is to optimize the accuracy of our model. We got the testing accuracy of the built-in logistic regression model is 83.33%. So, we try to identify the best parameter for optimizing the logistic regression algorithm. There are four types of parameters.

- 'penalty' used to prevent over-fitting i.e. l1, l2, elastic-net, none
- 'solver' used to Optimize the cost function i.e. lbfgs, liblinear, saga, newton-cg, sag
- 'C' means the number of learning rate i.e. 0.01, 0.1, 1, 10, 100
- 'max_iter' means the number of iterations i.e. 100, 500, 1000
- 'random_state' used to suffle the data i.e. none or integer value.

The parameter comparison of LR and OLR method are given on the Table *1*.

Method	penalty	solver	C	Max_iter	Random state
LR	l2	lbfgs	1	100	none
OLR	l1	liblinear	1	100	42

Table 1: parameter comparision of LR & OLR

We have found the best parameter and set them as optimal parameter for our model. The optimal parameters are 'C': 1, 'max_iter': 100, 'penalty': 'l1', 'solver': 'liblinear', 'random_state': '42'.

2) Hyperparameter Tuning and Experimental Result: GridSearchCV is used to pinpoint the perfect combination of hyperparameters that can increase our machine learning model's performance.

Cross-validation is used to partition the train data into two parts i.e. train and validation data. 5-fold (K=5) cross validation which divides the train data into five (5) partitions used by us. Which uses four (4) partitions for training and one (1) partition for testing in each iteration and continues 5 times. In the end, it gives the average performance [18].

For hyperparameter tuning, we use GridSearchCV with 5-fold cross-validation. The arguments are as follows:

- 'estimator' is A scikit-learn model which we proposed to predict.
- 'param grid' is a list of parameter values that we use to train our model.
- 'scoring' is the performance measure of our model, uses r2 for the regression model.

- 'cv' is the number of folds for k-fold cross-validation, uses 5-fold cross-validation.
- 'refit' is fitting with the best hyperparameter found in the 5-fold process i.e. R2
- 'n_jobs=-1' means using all processors of the local computer.

Now, we train the Logistic Regression model with the above arguments of GridSearchCV with 5-fold cross-validation and got an accuracy of 90.74% which is greater than the built-in LR model (83.33%). We named it as *Optimized Logistic Regression (OLR)* model.

III. Comparative Study of Machine Learning Techniques

In this part a comparative analysis will explain of our trained OLR model with five other machine learning techniques.

Confusion Matrix sum up the achievement of a machine learning classifier basis of test data. It shows the number of samplings. Here the matrix format in Table *2*.

True Positive (TP)	False Negative (FN)
False Positive (FP)	True Negative (TN)

Table 2: CONFUSION MATRIX

Where,

- TP refers to truly predict positive outcomes.
- TN refers to truly predict negative outcomes.
- FP stands for falsely predict positive outcomes
- FN stands for falsely predict negative outcomes.

A. Built-in Logistic Regression Classifier

The built-in logistic regression classifier is used to predict heart disease for the dataset. The achieved accuracy is 83.33%. The confusion matrix (Fig *4*) is given below.

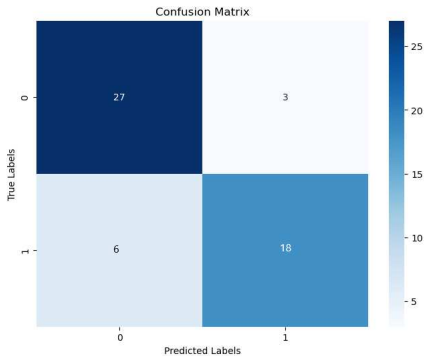

Fig 4: Confusion Matrix of Built-in LR classifier

B. Optimized Logistic Regression Classifier

This model is proposed by us described in section II. It was implemented using the best parameter and GridSearchCV with 5-fold cross validation. The achieved accuracy is 90.74%. The confusion matrix (Fig *5*) is given below.

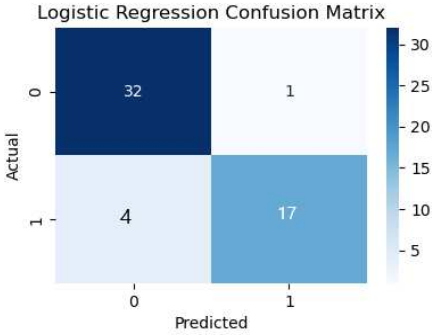

Fig 5: Confusion Matrix of Optimized LR classifier

C. Support Vector Machine(Linear) Classifier

Support Vector Machine (Linear kernel) classifier used for linearly separable data. It is faster than any other kernel of SVM[19]. It can capture complex relationship in data, but much computational cost required. It finds the optimal hyperplane that separates the target values. The achieved accuracy of the model is 87.03%. The confusion matrix (Fig *6*) of the linear kernel SVM for the dataset is given below.

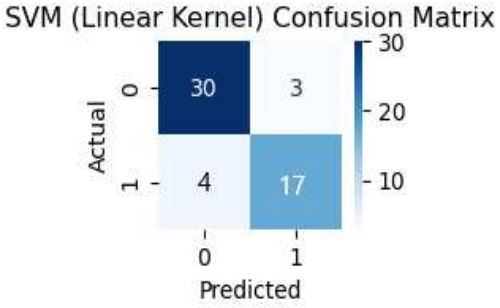

Fig 6: Confusion Matrix of SVM classifier

D. Random Forest Classifier

Random Forest is a supervised ML method that combines decision trees of multiple output and to make single outcomes. It can handle non-linearity very well, reduce overfitting and maintain high predictive accuracy[20]. The achieved accuracy of the model is 83.33%. The confusion matrix (Fig *7*) of the Random Forest for the dataset is given below.

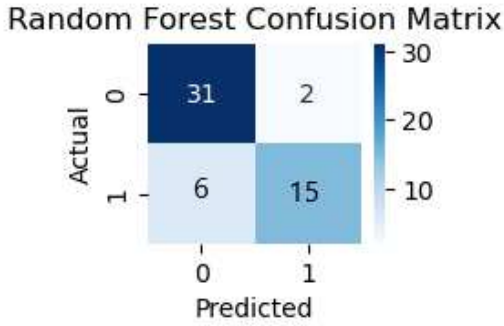

Fig 7: Confusion Matrix of RF classifier

E. AdaBoost Classifier

AdaBoost stands for Adaptive Boosting used to train the weak classifier and has the highest accuracy via the weak model given the highest weightage. It improves the performance by focusing on harder to classify features by iteration process, but it can be sensitve to noisy data and adds

complexity with each iteration. The achieved accuracy of the model is 88.88%. The confusion matrix (Fig *8*) of the AdaBoost for the dataset is given below.

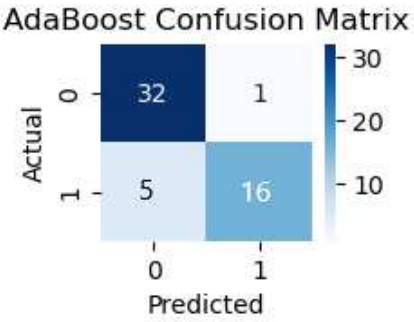

Fig 8: Confusion Matrix of AdaBoost classifier

F. Naïve Bayes Classifer

Naive Bayes helps classification problem based on Bayes theorem that assume feature independence. It is fast and predictable with high definition of data easier because of its simplicity. It remains effective method, when speed is prioritized over model complexity. The achieved accuracy of the model is 90.74%. The confusion matrix (Fig *9*) of the Naive Bayes for the dataset is given below.

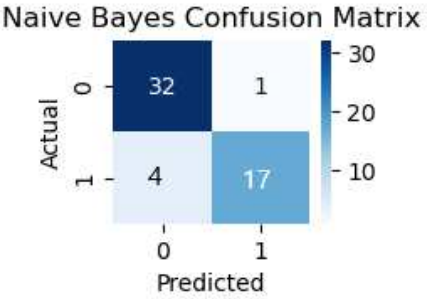

Fig 9: Confusion Matrix of Naïve Bayes classifier

For performance comparison, we illustrate a bar chart on Fig 10. Here are shown all the model performance that we compared.

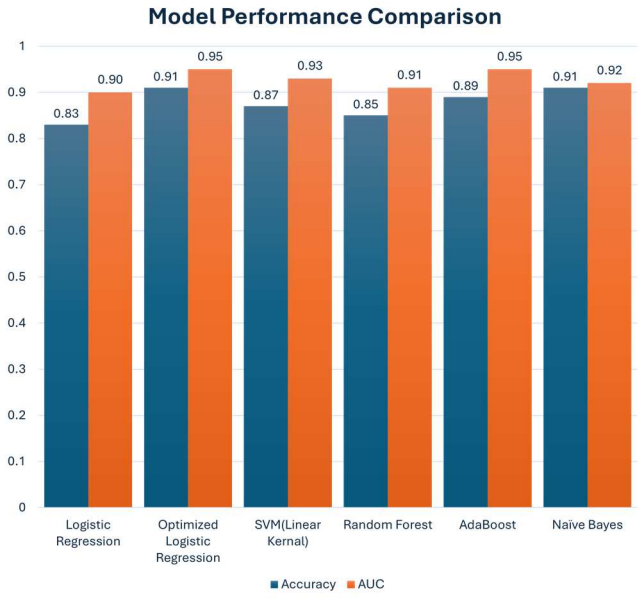

Fig 10: Model Performance Comparison

IV. Result and Discussion

In this part, we will discuss the result of our implemented model and compare it with other four machine learning techniques. A ROC (Receiver Operating Characteristic) curve is a graph that shows the achievement of a model at all classification thresholds[21]. It uses two parameters:

- True Positive Rate (TPR):

$$TPR = \frac{TP}{TP+FN} \ldots\ldots\ldots\ldots\ldots 2$$

- False Positive Rate (FPR):

$$FPR = \frac{FP}{FP+TN} \ldots\ldots\ldots\ldots\ldots 3$$

Area Under the Curve denotes as AUC. To compute the points in an ROC curve of different classification thresholds provided by AUC. It counts the entire 2D area underneath the ROC[21]. The ROC curve (Fig *11*) with AOC values for all the compared model is given below.

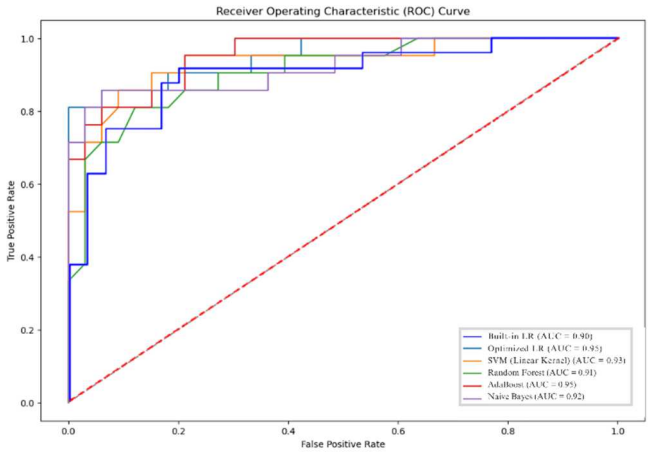

Fig 11: ROC curve with AUC values

Accuracy means how close the predicted and actual values are.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \ldots\ldots 4$$

Precision measures true positive prediction among all positive predictions.

$$Precision = \frac{TP}{TP+FP} \ldots\ldots\ldots\ldots 5$$

Recall measures the positive prediction among all actual positive items.

$$Recall = \frac{TP}{TP + FN} \ldots\ldots\ldots\ldots 6$$

f1-score is a metric that balances precision and recall.

$$f1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \ldots\ldots 7$$

Now, the accuracy of all models including precision, recall, f1-score, and AUC values are given below (Table *3*).

ML Model	Accuracy (%)	Precision (%)	Recall (%)	f1-score (%)	AUC
Logistic Regression	83.33	87	91	88.95	0.90
Optimized Logistic Regression	**90.74**	92	89	90.47	**0.95**
Support Vector Machine (Linear)	87.03	87	86	86.49	0.93
Random Forest	83.33	83	81	81.98	0.91
AdaBoost	88.88	90	87	88.47	0.95
Naïve Bayes	**90.74**	91	89	90.47	0.92

Table 3: COMPARATIVE ANALYSIS

Table *3* shows the detail experimental results among different multiple machine learning method. It shows that the optimized logistic regression (OLR) method obtained the maximum accuracy of 90.74 and AUC values of 0.95. Also, it appears the logistic regression, Linear Kernel Support Vector Machine, Random Forest, AdaBoost, Naïve Bayes classifiers have accuracy of 83.33%, 87.03%, 83.33%, 88.88%, 90.74% and AUC values are 0.90, 0.93, 0.91, 0.95, 0.92 respectively. However, Optimized Logistic Regression and Naïve Bayes classifier's all outcomes are same except AUC values. Higher AUC provides model performance more accurately. Thus, the proposed Optimized Logistic Regression (OLR) model can be predicted more accurately than a naive Bayes model.

V. CONCLUSION

This paper proposed an optimized logistic regression model that uses a dataset of medical records to train the machine learning model for early detection of heart disease. It achieved the highest prediction accuracy in heart disease probability, enabling early diagnosis and improved overall health outcomes for patients. In future studies, we will validate the proposal by collecting real hospital data and creating a dataset, so that we can build an accurate machine-learning model. In addition, an online application will be developed that will allow everyone to input information and more accurately detect heart disease.

REFERENCES

[1] World Heart Federation. (2023). *World Heart Report 2023*. World Heart Federation. https://www.world-heartfederation.org/resources/world-heart-report-2023

[2] Angell, S. Y., McConnell, M. V., Anderson, C. A., Bibbins-Domingo, K., Boyle, D. S., Capewell, S., Ezzati, M., De Ferranti, S., Gaskin, D. J., Goetzel, R. Z., Huffman, M. D., Jones, M., Khan, Y. M., Kim, S., Kumanyika, S. K., McCray, A. T., Merritt, R. K., Milstein, B., Mozaffarian, D., & Warner, J. J. (2020). The American Heart Association 2030 Impact Goal: A Presidential advisory from the American Heart Association. *Circulation, 141*(9), e120–e138. https://doi.org/10.1161/cir.0000000000000758

[3] Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. *Algorithms, 16*(2), Article 88. https://doi.org/10.3390/a16020088

[4] Sahoo, G., Kanike, K., Das, S., & Singh, P. (2022). Machine learning-based heart disease prediction: A study for home personalized care. In *2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1–6). IEEE. https://doi.org/10.1109/MLSP55214.2022.9943373

[5] Karthick, K., Aruna, S., Samikannu, R., Kuppusamy, R., Teekaraman, Y., & Thelkar, A. (2022). Implementation of a heart disease risk prediction model using machine learning. *Computational and Mathematical Methods in Medicine, 2022*, Article 6517716. https://doi.org/10.1155/2022/6517716

[6] Devi, K., Suruthi, S., & Shanthi, S. (2022). Coronary artery disease prediction using machine learning techniques. In *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)* (Vol. 1, pp. 1029–1034). IEEE. https://doi.org/10.1109/ICACCS54159.2022.9785140

[7] Gaikwad, M., Asole, P., & Bitla, L. (2022). Effective study of machine learning algorithms for heart disease prediction. In *2022 2nd International Conference on Power Electronics and IoT Applications in Renewable Energy and its Control (PARC)* (pp. 1–6). IEEE. https://doi.org/10.1109/PARC52418.2022.9726613

[8] Bhavekar, G. S., Das Goswami, A., Vasantrao, C. P., Gaikwad, A. K., Zade, A. V., & Vyawahare, H. (2023). Heart disease prediction using deep learning and machine learning techniques: A review. International Journal of Information Technology, 15, 1611-1623. https://doi.org/10.1007/s41870-022-00700-0

[9] Zhao, H. (2023). Visualization analysis and logistic regression-based heart disease risk prediction. *Applied and Computational Engineering*. https://doi.org/10.54254/2755-2721/16/20230880

[10] Vatsala, A., Poonam, S., & Gupta, S. (2023). A proficient framework for coronary artery disease prediction using logistic regression. *2023 IEEE International Conference on Communication, Computing, Power and Control Technologies (ICCPCT)*. https://doi.org/10.1109/iccpct58313.2023.10244967

[11] Soumyalatha, N., Shiler, N., & Ameen. (2023). Effective heart disease prediction framework using random forest and logistic regression. *2023 IEEE International Conference on Emerging Technologies in Computing and Networking (ViTECoN)*. https://doi.org/10.1109/ViTECoN58111.2023.10157078

[12] Geetha, M., & Pavithra, P. (2023). Estimating the risk of developing heart disease using the logistic regression model of machine learning. *International Journal for Science Technology and Engineering*. https://doi.org/10.22214/ijraset.2023.50542

[13] Mochammad, A. (2022). Predicting heart disease using logistic regression. *Knowledge Engineering and Data Science*. https://doi.org/10.17977/um018v5i22022p188-196

[14] Lal, P., Baliyan, P., Kochar, R., Singh Rathore, S., & Agarwal, P. (2024). Heart disease prediction using machine learning. In *Journal of Emerging Technologies and Innovative Research* (ISSN-2349-5162). Journal of Emerging Technologies and Innovative Research (JETIR). https://www.jetir.org/papers/JETIR2402583.pdf

[15] Statlog (Heart). (n.d.). *UCI Machine Learning Repository*. https://doi.org/10.24432/C57303

[16] Ahluwalia, H. (2024, March 19). Pair plots in machine learning. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2024/02/pair-plots-in-machinelearning/

[17] GeeksforGeeks. (2024, June 20). Logistic regression in machine learning. *GeeksforGeeks*. https://www.geeksforgeeks.org/understanding-logistic-regression/

[18] Shah, R. (2024, July 9). Tune hyperparameters with GridSearchCV. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/.

[19] Academy, I. (2024, April 4). Types of kernel in SVM—Kernels in support vector machine. *Theiotacademy*. https://www.theiotacademy.co/blog/types-of-kernel-insvm/#1-linear-kernel

[20] GeeksforGeeks. (2024, January 31). Random forest classifier using Scikit-learn. *GeeksforGeeks*. https://www.geeksforgeeks.org/random-forest-classifier-using-scikitlearn/

[21] Google for Developers. (n.d.). Classification: ROC curve and AUC. *Google for Developers*. https://developers.google.com/machine-learning/crashcourse/classification/roc-and-auc