Heart Disease Prediction Using Hybrid Model Integrating Artificial Neural Network, Decision Tree, and Logistic Regression

DOI: 10.38124/ijsrmt.v3i11.104

Nura Muhammad Sani[1*], Toochi C Ewunonu[2], Nwokonkwo, O.C[3], I.C EMETO[4], A. A. Galadima[5], and Aburrahman Buba[6]

[1] Department of Computer Science, Federal Polytechnic Kaltungo, Gombe Nigeria
[2] Department of Cyber Security, Federal University of Technology, Owerri Nigeria.
[3] Department of Information Technology Federal University of Technology Owerri
[4] Department of Cyber Security, Federal University of Technology, Owerri Nigeria.
[5] Department of Cyber Security, Federal University of Technology, Owerri Nigeria.
[6] Department of Computer Science, Federal Polytechnic Kaltungo, Gombe Nigeria

Corresponding Author: Nura Muhammad Sani[1*]

Abstract

Heart disease, a leading cause of global mortality, necessitates accurate prediction for timely intervention. This study proposes a hybrid model amalgamating LR, DT and ANN algorithms to enhance heart disease prediction. Using a Kaggle dataset comprising 1025 patient records with 14 features, including age, sex, chest pain, and cholesterol levels, the hybrid model achieved an impressive 88% precision. This outperforms individual models, with DT achieving 99% accuracy, LR with 80%, and ANN with 86%. Evaluation metrics demonstrate competitive performance, affirming the hybrid model as a robust tool for cardiovascular ailment prediction. The study underscores the efficacy of combining diverse algorithms, leveraging their strengths for more effective predictive modeling in cardiovascular health assessment.

Keywords: *Decision Tree, Logistic Regression, Artificial Neural Network, Hybrid Model.*

I. INTRODUCTION

The heart circulates blood, supplying nutrients and oxygen to bodily organs [1]. Heart disease impairs this function, affecting the brain, lungs, liver, and kidneys [1]. It causes cognitive and respiratory issues [2]. According to the WHO, heart disease is responsible for 31% of global annual deaths. [3]. Delayed diagnosis leads to poor outcomes and mortality [3]. Machine learning models show promise for early prediction to improve prognosis [4]. However, traditional models struggle with complexity while deep learning models require extensive data and risk overfitting [4]. A hybrid model combining both approaches could enable accurate, cost-effective early prediction.

II. RELATED WORK

Heart disease weakens the heart's ability to pump blood, leading to health issues like shortness of breath and fatigue [5]. It is prevalent globally, especially in developing countries, where limited diagnostic tools pose challenges for early diagnosis and treatment [5]. Accurate early prediction enables effective intervention, improving outcomes.

Machine learning techniques like logistic regression [6], random forests [7] and deep belief networks [9] have been explored for heart infection prediction using datasets from UCI repository and hospitals [5][10]. Classification accuracy ranges widely, from 77% [6] to 100% [3], reflecting differences in algorithms and data used. More complex ensemble approaches tend to outperform individual models. For example, Alizadehsani [11] combined multiple classifiers using an ensemble method, achieving 90% accuracy. Ruzzo-Tompa's [9] optimally configured deep belief network attained 94.61% accuracy. Feature engineering and selection techniques also boost performance [10].

How to Cite: sani, nura, C Ewunonu, T., Chukwuemeka, N., I.C, E., Galadima, A., & Buba, A. (2024). Heart Disease Prediction Using Hybrid Model Integrating Artificial Neural Network, Decision Tree, and Logistic Regression. International Journal of Scientific Research and Modern Technology, 3(11). https://doi.org/10.38124/ijsrmt.v3i11.104

However, issues like overfitting [13] and suboptimal model tuning [17] persist. Most studies test individual algorithms rather than hybrid approaches that could mitigate limitations. As Jabbar [13] demonstrated, minimizing training error sometimes increases testing errors. Ahmed [17] achieved only 80.32% validation accuracy despite testing multiple models.

Opportunities exist to advance accuracy further through larger, more diverse datasets [5], ensembles and hybrid methods [11], improved feature selection [10], and better hyper parameter optimization and cross-validation [10]. Combining complementary techniques in a tuned, thoroughly validated hybrid approach could yield both high accuracy and reliability in diagnosis.

III. METHODOLOGY

The learning aims to predict heart disease possibility using computerized prediction, benefiting medical professionals and patients. To accomplish this objective, a systematic approach was adopted. This process is illustrated in Figure 1 below:

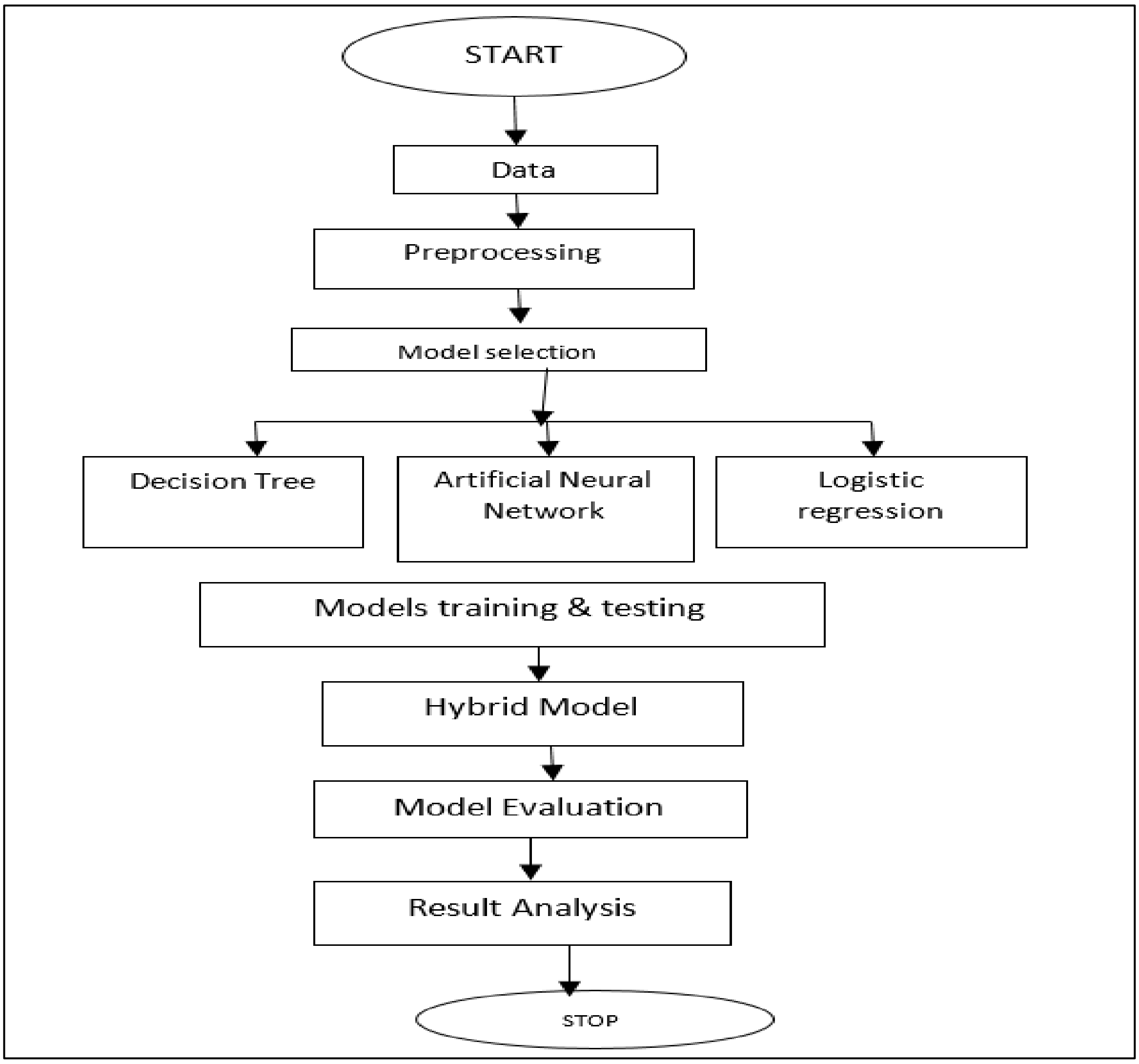

Fig 1 Block Diagram of the Model

- *Data Collection*

This study uses a Kaggle dataset of 1025 patient records with 14 distinct features related to heart disease risk factors.

- *Data Preprocessing*

To guarantee quality and consistency, the dataset underwent thorough preprocessing before the model was trained. This include fixing missing values, and normalizing features. The dataset's integrity was maintained by using typical scalar techniques to address missing values. To enable fair comparisons between various models, feature normalization was done to bring all characteristics to the same scale.

- *Performance Evaluation Metrics*

The models' predictive performance was assessed using various criteria, including F1-score, accuracy, precision, and recall. Precision gauges the proportion of correct positive predictions out of all positive predictions, while accuracy measures the overall correctness of the forecasts. The F1-score, which considers both false positives and false negatives, offers a balanced evaluation of a model's effectiveness by merging recall and precision. Recall indicates the percentage of correct positive predictions out of all actual positive instances. Together, these metrics offer a thorough insight into the strengths and limitations of the models

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1 - Score = 2\left(\frac{Precision*Recall}{Precision+Recall}\right) \quad (4)$$

IV. RESULT AND DISCUSSION

The research employed Google Colab on a system featuring an Intel(R) Pentium(R) processor with 8 GB RAM. The initial dataset comprised 1,025 rows and 14 attributes. However, following data cleaning and preprocessing, it was reduced to 820 rows and 13 attributes. The study involves various stages, starting with data collection, followed by preprocessing. Model selection is one of the next processes, and it includes LR, ANN, and DT. The models go through testing and training to produce a hybrid model. Using measurements like precision, recall, F1 score, and the area under the ROC curve, evaluation entails evaluating and interpreting the data. Hybrid models, DT, LG, and ANN are used in algorithm implementation. Model performance is measured using metrics such area under the ROC curve, F1 score, precision, and recall.

➢ *Data Balance*

This is used to understand the balance of the data. As shown in Figure 2 below, the data between cardiac (1) and non-cardiac (0) diseases are not balanced.

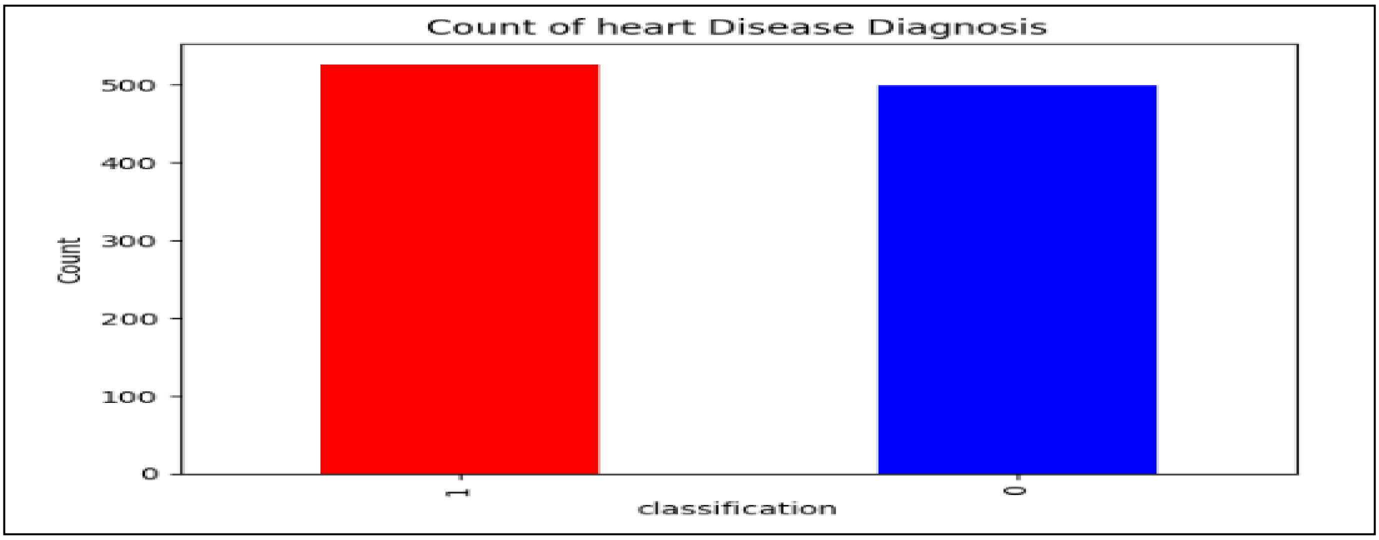

Fig 2 Data Balance

- *Correlation Matrix*

The correlation matrix was used to determine the relationship or strong correlations with the classification row. As shown in figure 3 bellow:

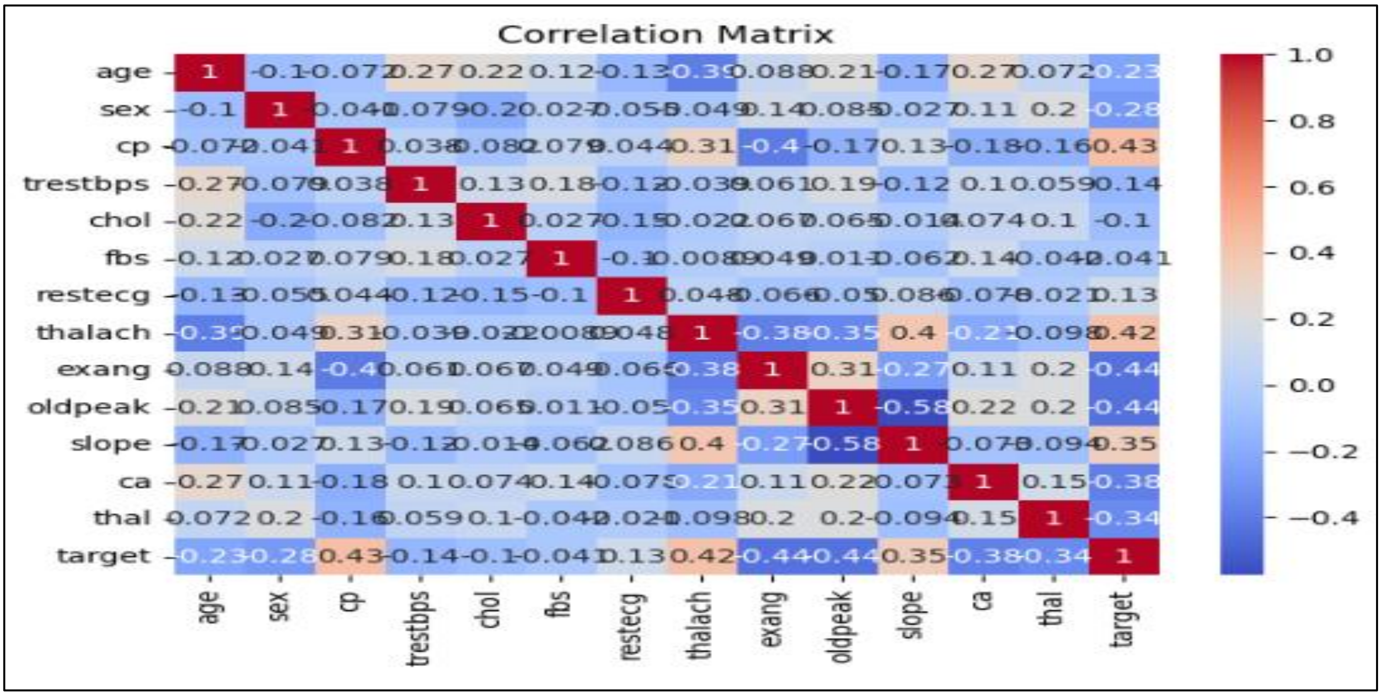

Fig 3 Correlation Matrix

➢ *Data Preprocessing*

The data was checked for missing values using data.isnull().sum() and no missing values were found. data.duplicated().sum() showed there were no duplicate rows. StandardScaler from sklearn.preprocessing was used to standardize the heart disease dataset features to have mean 0 and standard deviation 1. Standardization helps improve model performance and stability.

```
data.isnull().sum()

age          0
sex          0
cp           0
trestbps     0
chol         0
fbs          0
restecg      0
thalach      0
exang        0
oldpeak      0
slope        0
ca           0
thal         0
target       0
dtype: int64
```

Fig 4 The Total Number of Missing Values in Every Data Frame Column

➢ *Training Process*

The model training commenced with the fit method, ensuring specification of training (X_train, y_train) and validation data (X_test, y_test). Four models—Decision Tree, Logistic Regression, Artificial Neural Network (ANN), and Hybrid Model were employed for predicting heart disease. Decision Tree maximizes class separation through recursive feature-based splits. Logistic Regression estimates feature coefficients, modeling heart disease probability with a logistic function. ANN learns iteratively, adjusting weights and biases via back propagation. The Hybrid Model combines individual predictions, capturing diverse patterns and relationships for a comprehensive heart disease prediction approach.

➢ *Result of Artificial Neural Network*

In table 1 below, The study trained an artificial neural network for heart disease detection over 50 epochs. Results show training time, step loss, and validation loss trends indicating improving model accuracy and decreasing loss. Both training and validation accuracies were relatively high, suggesting good learning and generalization. Validation metrics help determine potential overfitting if validation loss increases while training loss decreases.

Table 1 Result of Artificial Neural Network

No	Training Time	Step loss	Accuracy
1.	1s 2ms/step	0.5692	0.7293
2.	0s 2ms/step	0.4863	0.7939
3.	0s 2ms/step	0.4278	0.8293
4.	0s 2ms/step	0.3843	0.8512
5.	0s 2ms/step	0.3538	0.8622
	.		
	.		
	.		
48.	0s 2ms/step	0.1745	0.9317
49.	0s 2ms/step	0.1718	0.9329
50.	0s 2ms/step	0.1693	0.9341

➢ *Hybrid Model Report*

In Table 2, the hybrid model demonstrates good balanced performance for heart disease prediction, with 90% and 86% precision, 85% and 90% recall for each class, and F1-scores of 0.87. It achieves an overall accuracy of 88% on a balanced test set. Macro and weighted average metrics show the model has balanced performance despite class imbalance.

Table 2 Hybrid Model Classification Report

	Precision	Recall	F1-Score	Support
0	0.9	0.85	0.87	102
1	0.86	0.9	0.88	103
Accuracy			0.88	205
Macro Avg	0.88	0.88	0.88	205
Wt. Avg	0.88	0.88	0.88	205

➢ *Confusion Matrix*

The confusion matrix was used in this study to provide detailed information about the model's predictions.

- *Logistic Regression Confusion Matrix*

Figure 9 shows that the model correctly predicted 73(71%) true negatives, 90(o.87) true positives, 13(12%) false positives, and 29 (28%) false negatives.

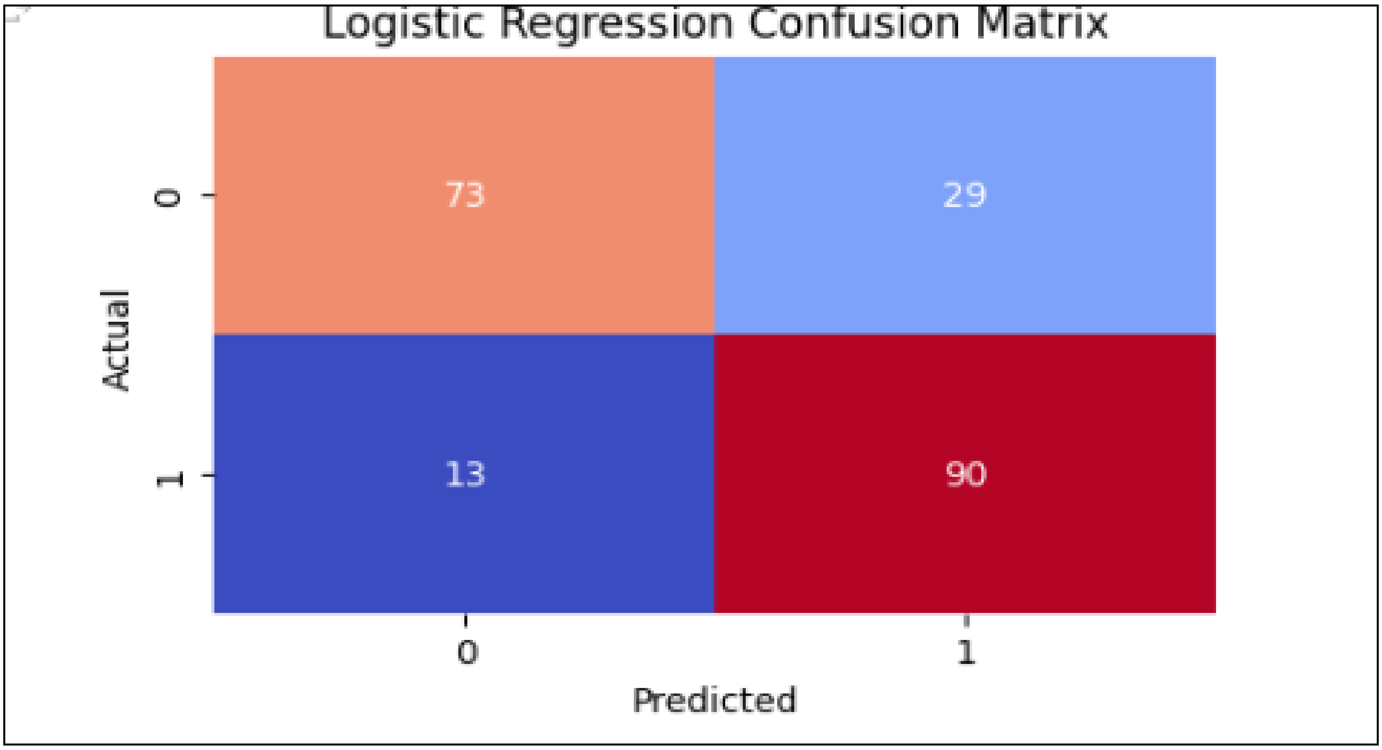

Fig 9 Logistic Regression Confusion Matrix

- *Decision Tree Confusion Matrix*

Figure 10 shows that the model correctly predicted 102 (100%) true negatives, 100(98%) true positives, 3 (0.029%) false positives, and 0 (0%) false negatives.

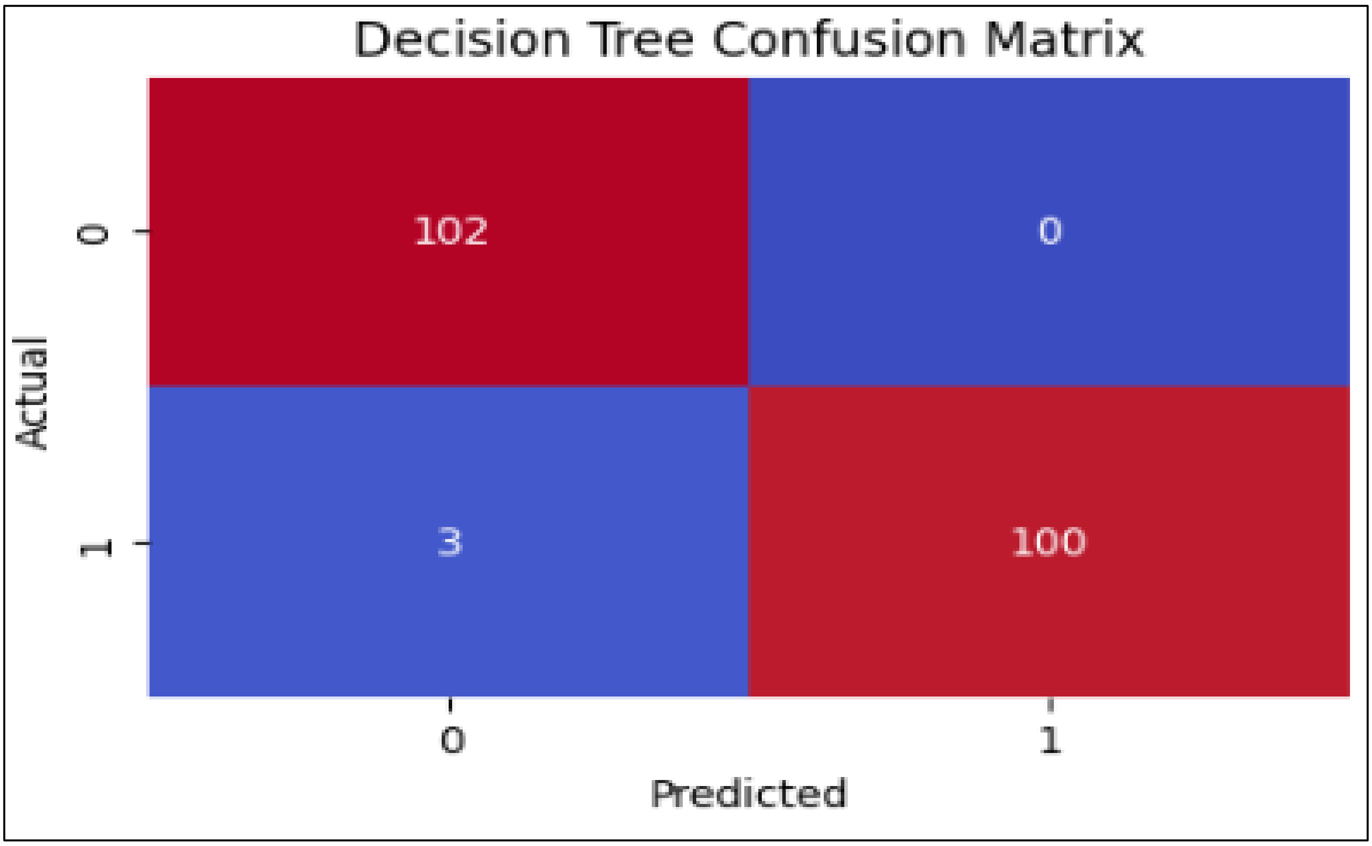

Fig 10 Decision Tree Confusion Matrix

- *Artificial neural network Confusion Matrix*

Figure 11 shows that the model correctly predicted 85 (83%) true negatives, 92(89%) true positives, 11 (10%) false positives, and 017(16%) false negatives.

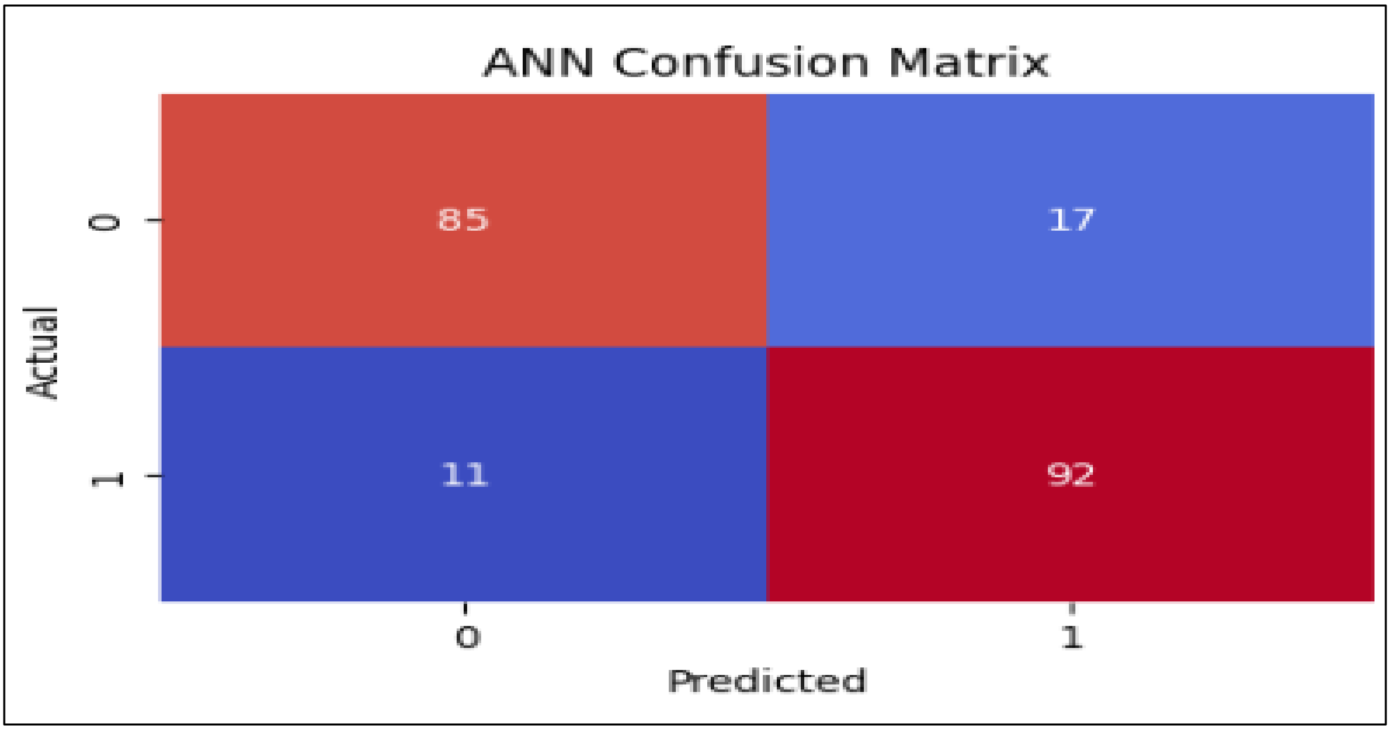

Fig 11 Artificial Neural Network Confusion Matrix

- *Hybrid model Confusion Matrix*

Figure 12 shows that the model correctly predicted 87 (85%) true negatives, 92(90%) true positives, 10 (9%) false positives, and 017(14%) false negatives.

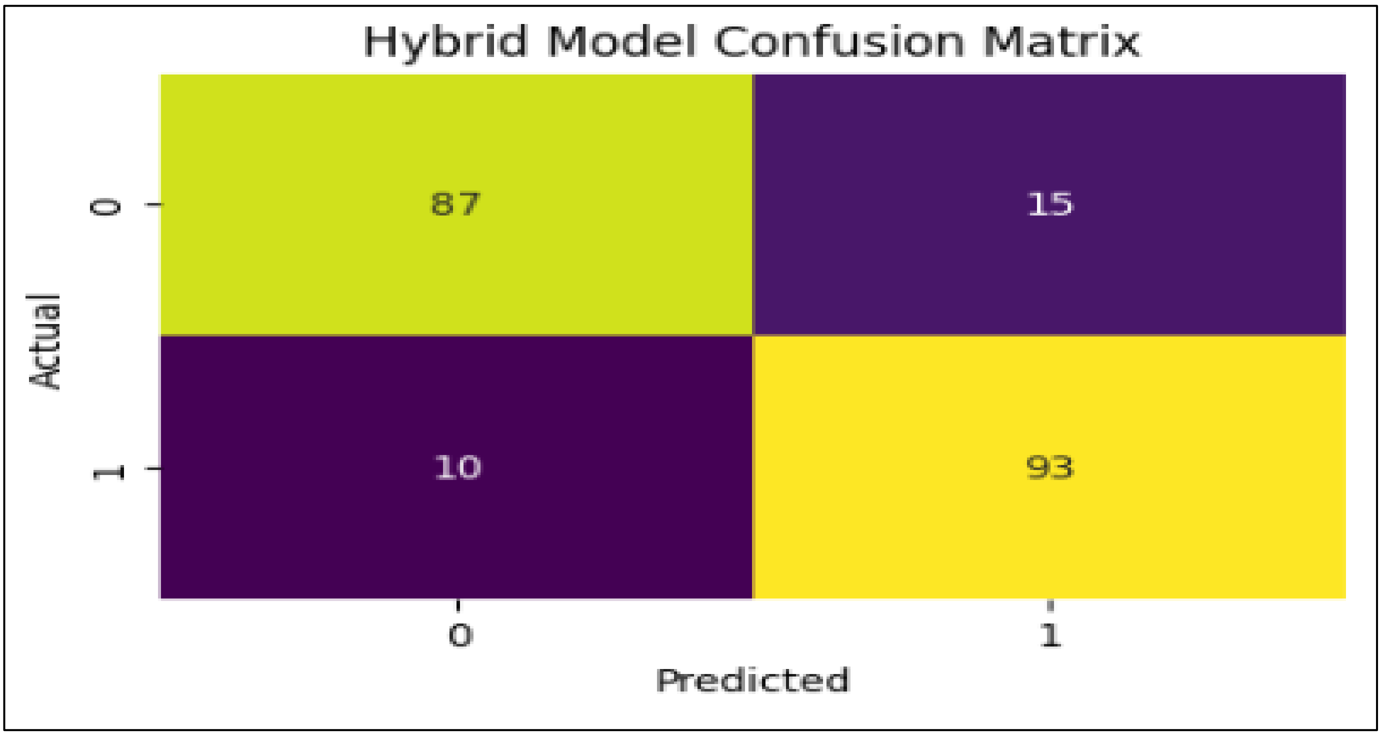

Fig 12 Hybrid Model Confusion Matrix

- *ROC Curve*

The ROC curves depict the models' ability to distinguish heart disease cases across varying thresholds. All models demonstrate excellent positive/negative class discrimination. The ANN model has an AUC of 0.96, performing between the Decision Tree and Logistic Regression. The Hybrid model achieves the highest AUC of 0.99, similar to the Decision Tree, indicating it distinguishes classes as effectively as the best standalone model.

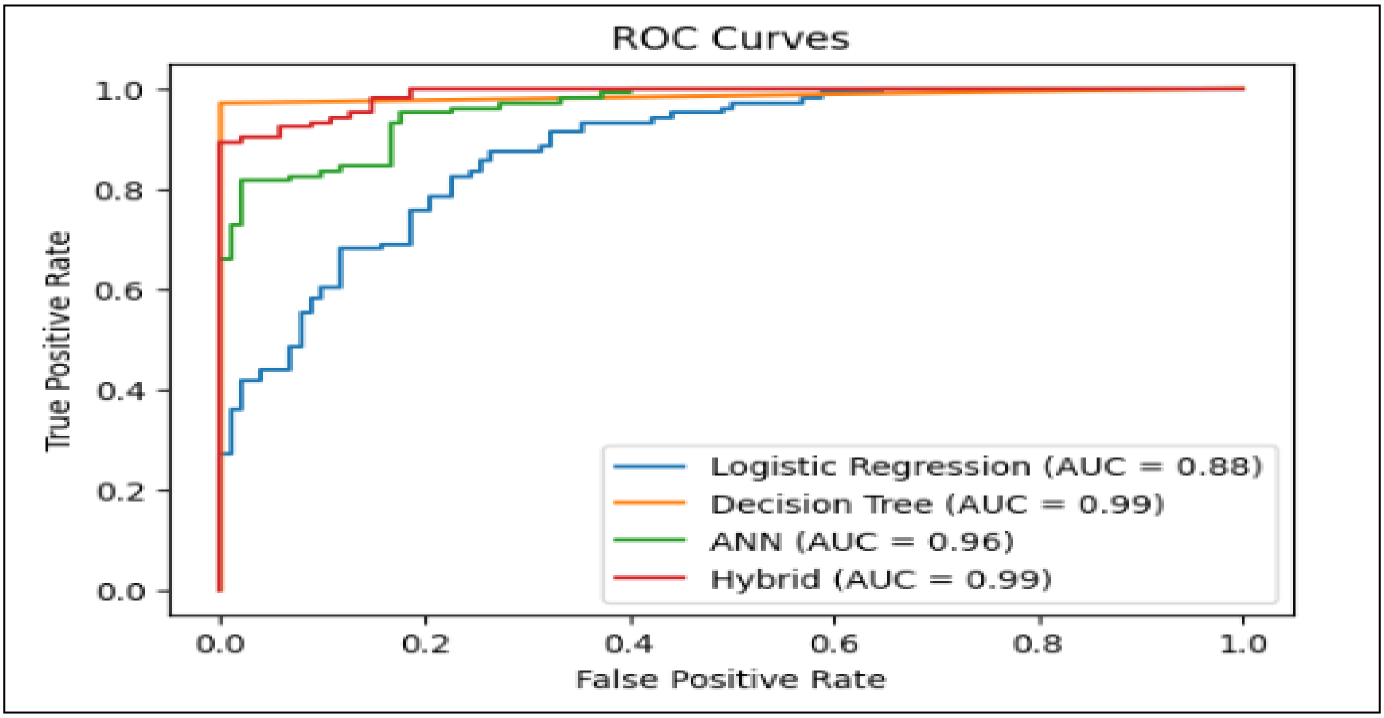

Fig 13 ROC Curves

➢ *Comparative Result Analysis*

The proposed hybrid model achieves a 0.88 accuracy, outperforming individual logistic regression (0.80) and ANN (0.86) models and approaching the 0.99 decision tree benchmark. The logistic regression model emphasizes recall for the positive class. The ANN model demonstrates balanced precision (0.84), recall (0.89) and F1-score (0.87) as shown in Table 3.

Table 3 Comparative Result Analysis

		Model	accuracy	Precision	Recall	F1Score	AUC
Proposed		HM	0.88	0.86	0.90	0.88	0.99
		DT	0.99	1.0	0.97	0.99	0.99
		LR	0.8	0.76	0.87	0.81	0.88
		ANN	0.86	0.84	0.89	0.87	0.96
Existing	[16]	MLP	0.87	0. 88	0. 84	0. 86	0.95
		RF	0. 87	0.89	0. 83	0. 86	0.95
		DT	0. 86	0.89	0.81	0. 85	0.94
		XGB	0.86	0.88	0. 83.	0.86	0.95
	[17]	SVM	0.92				0.78
		NB	0. 86				0.77
		LR	0. 85				0.78
		LightGBM	0.98				0.76
		XGBoost	0.99				0.72
		RF	100				0.92
	[18]	Naïve	0.86	0.82	0.87	0.89	
		Bayes	0.94	0.86	0.94	0.92	
		SVM&XGB	0.89	0.66	0.81	0.82	
		SVM and	0.95	0.97	0.94	0.95	
		DO					
		XGBoost					

[16], [17] and [18], showing that the model has achieved a high percentage of correct predictions.

V. CONCLUSION

To enhance heart disease prediction, the study created a hybrid model by combining DT, LR, and ANN. Compared to individual models, the hybrid model produced an accuracy of 0.88, which was almost identical to the 0.99 decision tree accuracy. It balanced the identification of positive instances and overall performance, exhibiting strong accuracy (0.86), recall (0.90), ROC-AUC (0.99), and F1-score (0.88). Overall, the hybrid model exceeded individual models, providing accurate and balanced heart disease predictions on par with the decision tree benchmark.

RECOMMENDATIONS

➢ *While the hybrid model shows promising results, further research could explore:*

- Fine-tuning model parameters to potentially improve performance.
- Incorporating additional features and advanced neural network architectures.
- Exploring ensemble techniques to combine models dynamically based on instance characteristics.

REFERENCES

[1]. "Structure and Function of the Heart," News-Medical.net, Sep. 5, 2022. [Online]. Available: https://www.news-medical.net/health/Structure-and-Function-of-the-Heart.aspx.

[2]. C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective heart disease prediction using machine learning techniques," Algorithms, vol. 16, no. 2, p. 88, 2023.

[3]. B. Paul and B. Karn, "Heart disease prediction using scaled conjugate gradient backpropagation of artificial neural network," Soft Computing, vol. 27, no. 10, pp. 6687-6702, 2023

[4]. B. Sun, W. Cui, G. Liu, B. Zhou, and W. Zhao, "A hybrid strategy of AutoML and SHAP for automated and explainable concrete strength prediction," Case Studies in Construction Materials, vol. 19, p. e02405, 2023.

[5]. S. Begum, "A study for predicting heart disease using machine learning," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 10, pp. 4584-4592, 2021.

[6]. M. Gudadhe, K. Wankhade, and S. Dongre, "Decision support system for heart disease based on support vector machine and articial neural network," in Proceedings of International Conference on Computer and Communication Technology (ICCCT), pp. 741-745, Allahabad, India, September 2010.

[7]. M. Kavitha, G. Gnaneswar, R. Dinesh, Y. Rohith Sai, and R. Sai Suraj, "Heart disease prediction using hybrid machine learning model," in 2021 6th International Conference on Inventive Computation Technologies (ICICT), pp. 1329-1333, IEEE, 2021.

[8]. J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," IEEE Access, vol. 8, pp. 107562-107582, 2020.

[9]. S. A. Ali, B. Raza, A. K. Malik, A. R. Shahid, M. Faheem, H. Alquhayz, and Y. J. Kumar, "An Optimally Configured and Improved Deep Belief Network (OCI-DBN) Approach for Heart Disease Prediction Based on Ruzzo–Tompa and Stacked Genetic Algorithm," IEEE Access, vol. 8, pp. 65947-65958, 2020.

[10]. S. Ambesange, A. Vijayalaxmi, S. Sridevi, Venkateswaran and B. S. Yashoda, "Multiple Heart Diseases Prediction using Logistic Regression with Ensemble and Hyper Parameter tuning Techniques," 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), London, UK, 2020, pp. 827-832, doi: 10.1109/WorldS450073.2020.9210404.

[11]. R. Atallah and A. Al-Mousa, "Heart disease detection using machine learning majority voting ensemble method," in 2019 2nd International Conference on New Trends in Computing Sciences (ICTCS), October 2019, pp. 1-6.

[12]. S. Ismaeel, A. Miri, and D. Chourishi, "Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis," in Proceedings of the 2015 IEEE Canada International Humanitarian Technology Conference (IHTC2015), Ottawa, ON, Canada, 31 May–4 June 2015, pp. 1–3.

[13]. K.H. Miao, J.H. Miao, and G.J. Miao, "Diagnosing Coronary Heart Disease using Ensemble Machine Learning," Int. J. Adv. Comput. Sci. Appl., vol. 7, pp. 30–39, 2016.

[14]. A. Khan, M. Qureshi, M. Daniyal, and K. Tawiah, "A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction," Health & Social Care in the Community, 2023.

[15]. S. Nashif, M. R. Raihan, M. R. Islam, and H. Imam, "Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System," World Journal of Engineering and Technology, Scientific Research Publishing, Jan. 1, 2018. [Online]. Available: https://doi.org/10.4236/wjet.2018.64057.

[16]. C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," Algorithms, vol. 16, no. 2, p. 88, 2023. [Online]. Available: https://doi.org/10.3390/a16020088.

[17]. K. Karthick, S. Aruna, R. Samikannu, R. Kuppusamy, Y. Teekaraman, and A. R. Thelkar, "Implementation of a Heart Disease Risk Prediction Model Using Machine Learning," Computational and Mathematical Methods in Medicine, Hindawi Publishing Corporation, May 2, 2022. [Online]. Available: https://doi.org/10.1155/2022/6517716.

[18]. U. Nagavelli, D. Samanta, and P. Chakraborty, "Machine Learning Technology-Based Heart Disease Detection Models," Journal of Healthcare Engineering, Hindawi Publishing Corporation, February 27, 2022. [Online]. Available: https://doi.org/10.1155/2022/7351061.