Efficient Machine Learning Algorithms for Cardiovascular Risk Prediction

V. Sitharamulu
Department of Computer Science and Engineering, GITAM School of Technology, GITAM (Deemed to be University), Hyderabad, Telangana, India
vsitaramu.1234@gmail.com

Sreerama Murty Maturi
Department of Computer Science and Engineering, GITAM (deemed to be University), Hyderabad, India
sreeramsssit@gmail.com (corresponding author)

M. Murugesan
Department of Computer Science and Engineering, Anurag Engineering College Ananthagiri (V&M) Suryapet (Dt) -508 206, India
murugeshvim@gmail.com

Mahammad Rafi Dudekula
Department of Computer Science and Engineering (Data Science), Institute of Aeronautical Engineering (IARE), Dundigal-500043, Hyderabad, India
mahammadrafi0780@gmail.com

Hanumantha Rao Battu
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (KLEF), KLEF Deemed to be University, Green Fields, Vaddeswaram, Guntur District, AP-522 302, India
hanuma9999@yahoo.com

Received: 17 June 2025 | Revised: 1 July 2025, 8 July 2025, and 11 July 2025 | Accepted: 13 July 2025

 DOI: https://doi.org/10.48084/etasr.12795

ABSTRACT

Cardiovascular Disease (CVD) is a critical global health concern requiring efficient early prediction. This study evaluates Machine Learning (ML) algorithms—Logistic Regression (LR), Light Gradient Boosting Machine (LGBM), Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbors (K-NN)—for forecasting the cardiovascular risk. By analyzing key patient features, such as cholesterol, blood pressure, and lifestyle, these models classify individuals into low or high-risk categories. The findings demonstrate that LGBM and RF achieve superior performance, both reaching 99% accuracy, precision, recall, and F1-score, alongside high ROC-Area Under the Curve (AUC) values. This research provides robust, data-driven tools to enhance the timely diagnosis and support preventative medical strategies, ultimately improving the patient outcomes.

Keywords-heart disease prediction; machine learning; early diagnosis; real-time prediction systems; healthcare AI

I. INTRODUCTION

Heart disease encompasses a range of cardiovascular disorders, including heart failure and arrhythmias. If not identified early, these disorders can lead to serious health complications. Traditional diagnostic techniques, such as blood tests, ECGs, and physical examinations, can be time-consuming and often require a high level of medical expertise. Treatment may also be postponed if the early-stage symptoms are overlooked. This necessitates the development of sophisticated predictive algorithms, which can effectively identify the individuals at high risk, within the medical practice.

The advances in artificial intelligence have established ML as a potent tool for cardiac disease prediction and medical data evaluation. By processing vast volumes of patient data, identifying hidden patterns, and generating data-driven

predictions, ML models can enhance the accuracy and reduce the human errors. This study examines the effectiveness of several ML techniques for predicting the cardiac disease. Incorporating ML into the diagnostic process enhances the early detection, leading to improved patient outcomes.

With millions of fatalities annually, CVD is one of the world's leading causes of death. The early detection and prevention of CVD can considerably decrease the mortality rates and enhance the patient outcomes. Through the analysis of extensive datasets and the discovery of intricate patterns, ML algorithms have demonstrated considerable promise in the prediction of the cardiovascular risk. Numerous risk variables, such as age, sex, blood pressure, cholesterol, smoking status, and family history, have an impact on CVD, which is a complex illness. Accurately estimating the cardiovascular risk in a variety of groups is a challenge for traditional risk prediction algorithms, like the Framingham Risk Score.

This work proposes an ML-based approach to enhance the accuracy and reliability of the heart disease prediction by comparing performance, using multiple classification techniques. The primary goal is to develop an efficient predictive model that would support the early diagnosis and risk assessment for healthcare professionals. To identify which model is the most reliable and accurate for predicting the heart disease, the proposed method employs a range of classification algorithms, including RF, DT, LGBM, and LR. When different models are compared, the model with the best performance is selected for real-world applications.

II. RELATED WORK

Previous studies have explored ML for cardiovascular risk prediction, each with specific contributions and limitations. Authors in [1] used ensemble methods but noted constraints in the dataset size, class imbalance, and real-time validation. Various ML algorithms have been investigated for heart disease prediction, including supervised models [4, 10, 13] and active learning-based models [11]. LR, for instance, has been examined for the diabetes risk [3] and heart disease prediction [4, 8].

Optimized and hybrid models have also been developed to enhance prediction. Authors in [2] developed hybrid optimized models, yet faced challenges in real-time scalability and clinical evaluation. Other optimization techniques, such as PSO with ML [9] and MLP-PSO hybrid algorithms [12], have also been applied.

Regarding the data and model complexity, authors in [5] demonstrated good prediction ability but pointed out limitations in data diversity and the absence of deep learning methods. Similarly, Authors in [6] evaluated various algorithms (KNN, LR, RF) but were constrained by the limited features and dataset size, suggesting the potential of deep learning and e-Health integration. Deep learning approaches have indeed been explored for cardiovascular risk prediction [7]. Early and accurate detection using intelligent computational models remains a key area of research [14].

Building on these efforts, the current work addresses these gaps by employing a wider range of ML algorithms on a larger, more diverse dataset, incorporating advanced optimization and resampling methods to enhance the prediction accuracy, generalizability, and clinical applicability.

III. METHODOLOGY

The following key features are used for analyzing and predicting the cardiovascular risk.

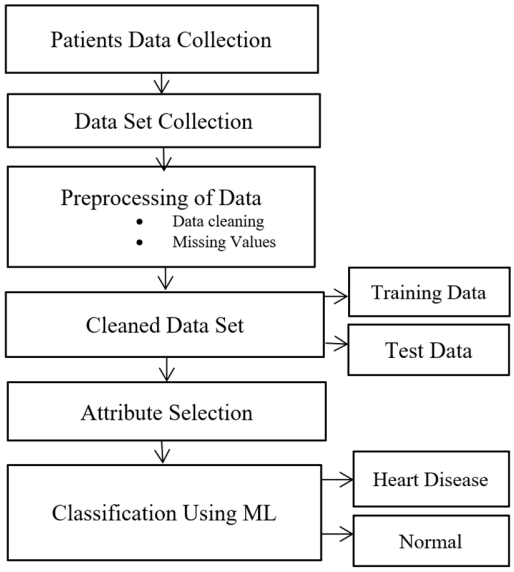


Fig. 1. Methodology for cardiovascular risk prediction.

A. *Data Collection*

- The dataset on the heart disease was provided by one of India's multispecialty hospitals.

B. *Data Pre-Processing*

- Removal of the missing values and problematic entries (inconsistent records).
- To deal with the missing values, imputation techniques were applied.
- The Z-Score method was performed for outlier elimination.
- Feature scaling: To ensure uniformity, the numerical values were subjected to normalization or standardization.

C. *Encoding Categorical Variables*

For the categorical variables, such as gender and type of chest discomfort, encoding was applied.

D. *Dataset Splitting*

To evaluate the model performance, the dataset was split into training, validation, and test sets. This splitting strategy prevents the data leakage and ensures reliable evaluation.

E. *Feature Selection*

To improve the model efficiency and reduce the unnecessary data volume, the following techniques were applied:

- Correlation Matrix: Identifies and removes highly correlated features to prevent redundancy.
- Principal Component Analysis (PCA): Maintains the important information while reducing dimensionality.
- Select K-Best: Selects the most important features based on statistical relevance.

F. *Model Training and Selection*

To predict the heart diseases, various ML models were trained. These involved the LR, DT, RF, KNN, and Gradient Boosting Algorithms. Hyper parameter tuning was performed using either random or grid-based search.

G. *Model Evaluation*

The models are assessed using various performance metrics:

- Accuracy: Indicates how accurate is the overall prediction.
- The reliability is increased and class imbalance is managed with the aid of precision and recall.
- F1-score: Improves the classification by striking a balance between the recall and precision.
- ROC-AUC Curve: Assesses how well the model can differentiate across classes.
- Cross-validation: Guarantees a solid and trustworthy performance evaluation.

H. *Classification Metrics*

The number of accurate guesses divided by the total number of predictions is the classification accuracy. The ratio of TPs to the total expected positives is known as precision. The ratio of TP to all of the ground truth's positives is called recall. The F1-score is the harmonic mean of precision and recall. TPR and FPR are used in the AUC-ROC score. The matching formulas are:

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

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

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

$$\text{F1 \_ score} = 2\ X\ \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}} \quad (4)$$

$$\text{TPR} = \frac{TP}{TP+FN} \quad (5)$$

$$\text{FPR} = \frac{FP}{FP+TN} \quad (6)$$

I. *Model Training and Optimization*

To ensure that the developed models generalize effectively to unseen patient data and to mitigate overfitting, rigorous hyperparameter tuning was performed for each algorithm. This optimization process aimed to maximize the model accuracy and robustness.

J. *Deployment and Interpretation*

- The best-performing model was deployed using Flask or Streamlet for real-time predictions.
- Predictions were visualized using interactive dashboards, graphs, and reports for a better interpretation and decision-making.

IV. IMPLEMENTATION

Patient records containing a variety of characteristics, including age, blood pressure, cholesterol, and lifestyle choices, make up the dataset used to feed the ML models.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	patientid	age	gender	chestpain	restingBP	serumcholestrol	fastingbloodsugar	restingrelectro	maxheartrate	exerciseangia	oldpeak	slope	noofmajorvessels	target
2	103368	53	1	2	171	0	0	1	147	0	5.3	3	3	1
3	119250	40	1	0	94	229	0	1	115	0	3.7	1	1	0
4	119372	49	1	2	133	142	0	0	202	1	5	1	0	0
5	132514	43	1	0	138	295	1	1	153	0	3.2	2	2	1
6	146211	31	1	1	199	0	0	2	136	0	5.3	3	2	1
7	148462	24	1	1	173	0	0	0	161	0	4.7	3	2	1
8	168686	79	1	2	130	240	0	2	157	0	2.5	2	1	1
9	170498	52	1	0	127	345	0	0	192	1	4.9	1	0	0
10	188225	62	1	0	121	357	0	1	138	0	2.8	0	0	0
11	192523	61	0	0	190	181	0	1	150	0	2.9	2	0	1
12	201030	59	0	1	190	529	1	1	151	1	3.2	2	2	1
13	208877	58	1	2	192	409	1	0	138	0	2.3	3	1	1
14	223295	27	1	0	129	135	0	1	192	1	1	0	0	0
15	226481	59	1	0	98	209	0	0	117	1	5.6	1	0	0
16	229445	58	1	0	170	354	0	0	170	0	5.6	1	0	0
17	235344	32	1	2	188	0	0	0	134	1	4.5	2	3	1
18	236763	42	0	3	137	350	0	1	110	0	3.2	2	2	1
19	240461	65	1	0	200	247	1	1	194	1	3.7	1	1	0
20	247055	59	1	2	182	177	0	1	168	0	2.1	2	1	1
21	260870	35	1	0	127	269	0	0	87	1	3.8	0	1	0
22	266839	39	1	3	196	253	1	2	140	1	3.5	2	1	1
23	322287	72	1	1	177	397	0	2	124	0	5.2	3	1	1
24	327110	24	0	0	136	164	0	0	91	1	1.8	1	1	0
25	335359	59	1	2	156	223	0	2	184	1	2.4	2	2	1

Cardiovascular_Disease_Dataset

Fig. 2. Cardiovascular dataset.

The dataset contains 1000 rows and 14 columns, with 12 columns being used as predictive attributes. After the cleaning process, 976 rows were left. The crucial attributes, namely age, gender, kind of chest pain, resting blood pressure, serum cholesterol, fasting blood sugar, and ECG readings, are among the traits that are significant indicators of the cardiovascular health. The number of major vessels, the slope of the ST segment, exercise-induced angina, maximum heart rate, and ST depression (old peak) are also shown; these details are helpful in assessing the heart disease problem. The target variable (1 for cardiac disease, 0 for no disease) makes this dataset perfect for supervised ML models. A predictive model may be created to categorize people according to their heart disease risk by pre-processing and evaluating these parameters. This will help with the early diagnosis and preventative healthcare treatments.

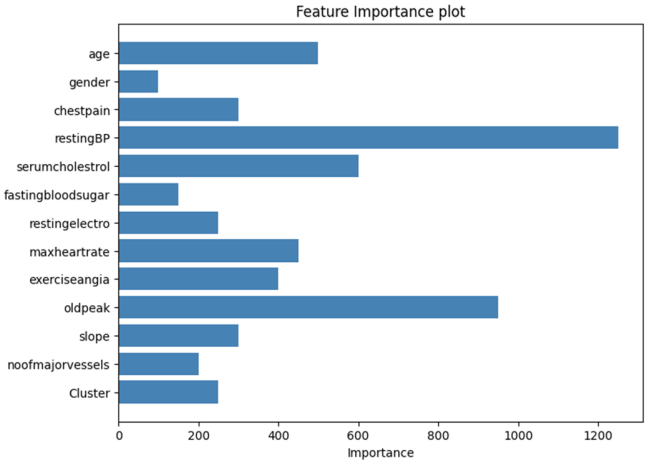


Fig. 3. Feature importance of dataset characteristics.

V. RESULTS AND DISCUSSION

A. *Decision Tree Classification Report*

Using precision, recall, F1-score, and support as metrics, this report assesses the DT model's performance.

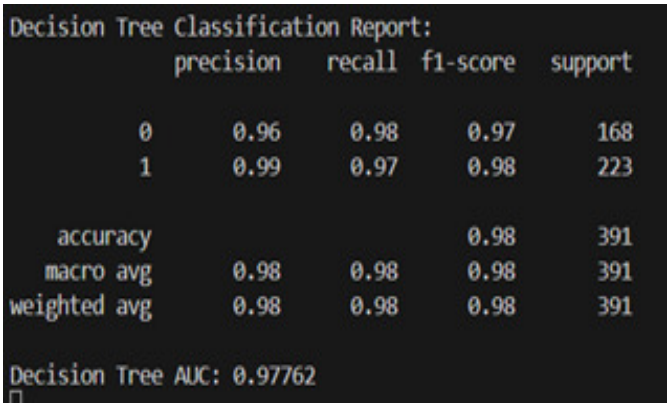


Fig. 4. Performance metrics for DT.

TABLE I. METRICS FOR THE DT MODEL

	Class 0: negative situations	**Class 1: positive situations**
Precision	0.96	0.99
Recall	0.98	0.97
F1-score	0.97	0.98
Support	168	223

- Accuracy: 0.98 (98%) => the percentage of the test samples that were properly classified.
- Macro Average: 0.98 → the mean of the F1-score, recall, and precision for every class.
- Similar to the macro average, but taking class imbalance into account, is the weighted average of 0.98

The differentiated model classes are measured by their AUC, which is 0.97762 (~97.76%). A value near 1.0 indicates that the model does well in categorization.

B. *Result Comparison*

The difference between Table II and Table III is in the experimental setup. In Table II, the results come from cross-validation after optimization. The results in Table III are derived from unseen test-set data. Table II shows a higher accuracy for all models compared to Table III, as expected.

TABLE II. COMPARISON METRICS OF ML ALGORITHMS (CROSS-VALIDATION)

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**ROC-AUC (%)**
LGBM	99%	99%	99%	99%	99.98%
RF	99%	99%	99%	99%	99.93%
LR	97.70%	98%	98%	98%	99.68%
DT	98%	98%	98%	98%	97.76%
K-NN	97%	97%	97%	97%	98.88%

TABLE III. ACCURACY OF ML MODELS (TEST SET)

Model	**Accuracy (%)**
LGBM	89.17%
RF	85.25%
LR	85.84%
DT	75.58%
K-NN	51.19%

- **Accuracy of the RF Classifier: 99%.** It had the best accuracy, demonstrating a robust generalization and excellent handling of the feature diversity. It is quite dependable for predicting the heart disease because of its excellent recall (99.93%), which guarantees few false negatives.
- **Accuracy of the LGBM: 99%.** Ensemble learning was used to achieve a competitive performance while successfully lowering the bias and variance using boosting strategies. Excellent predictive capacity was ensured by its high recall (98.21%) and robust ROC-AUC score (99.98%).
- **Effectiveness of LR: 97.70%.** An easy-to-use paradigm that works well with data that can be separated linearly.

With a 98% accuracy rate and a 98% F1-score, it continues to be a solid starting point.

- **K-NN:** Achieved a strong 97% accuracy and an excellent ROC-AUC score of 98.88%. While these metrics demonstrate a robust performance and good discriminative ability, K-NN was comparatively outperformed by other models in this study.

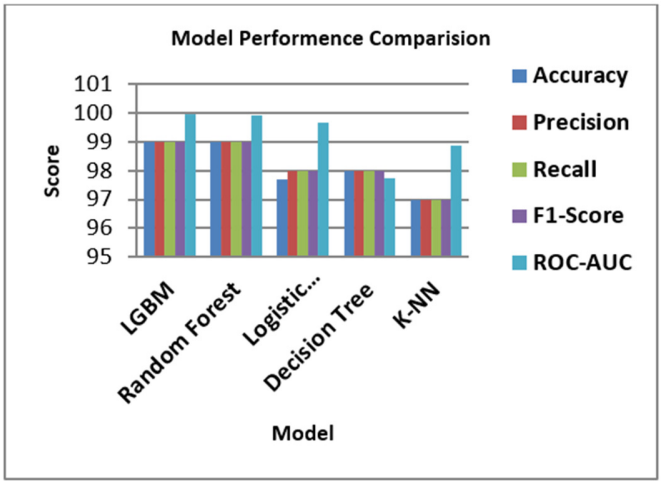


Fig. 5. Evaluation of model performance.

In Figures 5 and 6, performance indicators, including F1-score, ROC-AUC, recall, accuracy, and precision, are used to assess the trained models. The model's prediction power and classification errors are examined using a confusion matrix. Among the various models tested, LGBM and RF demonstrated the highest accuracy. These models effectively handle the complex patterns in the data, making them the most reliable choices for deployment. While other models, such as LR, DT, and K-NN, also performed well, LGBM and RF outperformed them in terms of overall accuracy and robustness. This comparison ensures that the most efficient models are selected for early heart disease prediction.

(a)

KNN Classification Report:

	precision	recall	f1-score	support
0	0.95	0.98	0.96	168
1	0.99	0.96	0.97	223
accuracy			0.97	391
macro avg	0.97	0.97	0.97	391
weighted avg	0.97	0.97	0.97	391

KNN AUC: 0.98879

(b)

LightGBM Classification Report:

	precision	recall	f1-score	support
0	0.98	0.99	0.99	168
1	1.00	0.98	0.99	223
accuracy			0.99	391
macro avg	0.99	0.99	0.99	391
weighted avg	0.99	0.99	0.99	391

```
C:\Users\PANDUREDDY\AppData\Local\Packages\PythonSoftwar
precation.py:151: FutureWarning: 'force_all_finite' was
  warnings.warn(
```

LightGBM AUC: 0.99979

(c)

Logistic Regression Accuracy: 97.69821

Logistic Regression Classification Report:

	precision	recall	f1-score	support
0	0.98	0.97	0.97	168
1	0.98	0.98	0.98	223
accuracy			0.98	391
macro avg	0.98	0.98	0.98	391
weighted avg	0.98	0.98	0.98	391

Logistic Regression AUC: 0.99680

(d)

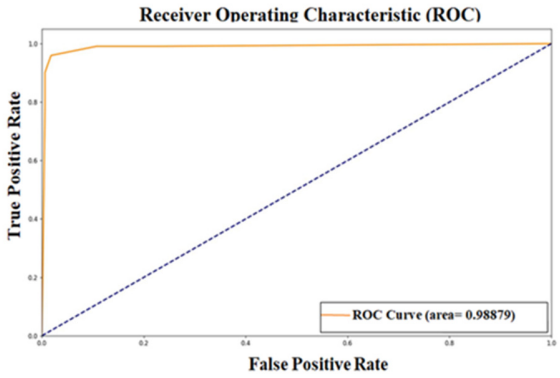
Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.97	1.00	0.99	168
1	1.00	0.98	0.99	223
accuracy			0.99	391
macro avg	0.99	0.99	0.99	391
weighted avg	0.99	0.99	0.99	391

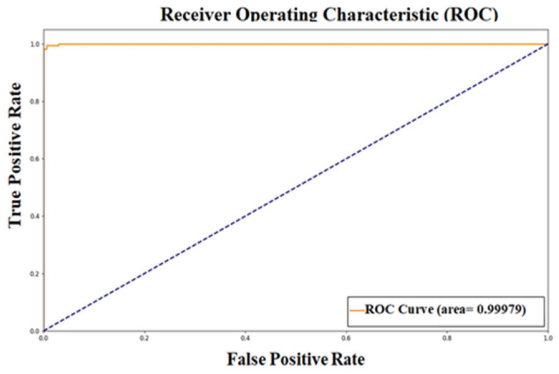
Random Forest AUC: 0.99931

Fig. 6. Performance metrics for: (a) K-NN, (b) LGBM, (c) LR, (d) RF.

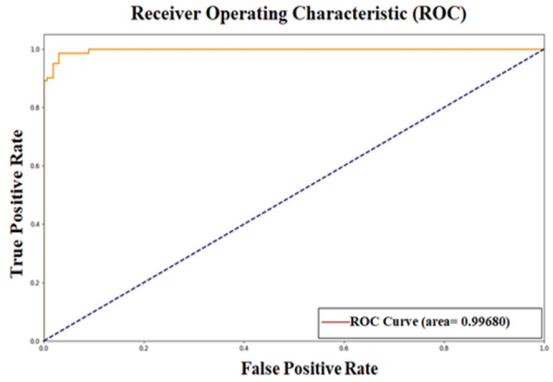
(a)



(b)



(c)



(d)

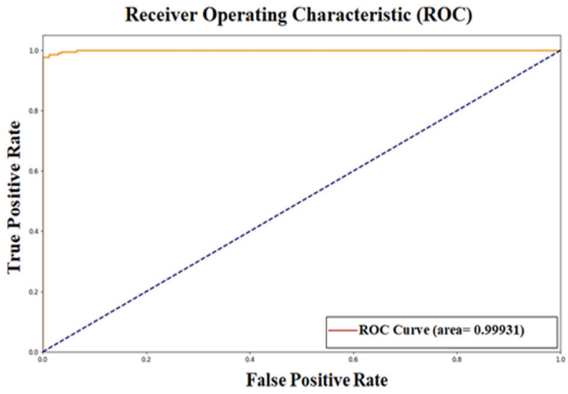


Fig. 7. AUC curve for: (a) K-NN, (b) LGBM, (c) LR, (d) RF.

VI. CONCLUSION

This study successfully evaluated several Machine Learning (ML) models for the cardiovascular risk prediction. The findings demonstrate that Logistic Regression (LR), Light Gradient Boosting Machine (LGBM) and Random Forest (RF) achieved superior performance, both reaching 99% accuracy, precision, recall, and F1-score, alongside high ROC-AUC values in cross-validation. While other models, like LR, Decision Tree (DT), and K-Nearest Neighbors (K-NN) also showed promising results, LGBM and RF proved to be the most robust and accurate for this dataset. These data-driven tools offer significant potential to enhance the timely diagnosis and support preventative medical strategies, ultimately improving the patient outcomes.

VII. FUTURE SCOPE

Building on the current study's promising results, future research will focus on several key areas to enhance the predictive capabilities and clinical utility of cardiovascular risk models:

- Real-time and Multimodal Data Fusion: Incorporating real-time physiological data from wearables/IoT devices, patient habits, and fusing diverse data types (e.g., clinical, imaging, genomic) for comprehensive and dynamic risk prediction.
- Explainable AI (XAI): Integrating XAI methods to enhance the model transparency and trustworthiness, crucial for clinical adoption and decision-making.
- Personalized Medicine and Clinical Deployment: Developing highly individualized prediction models and deploying them as robust clinical decision support systems (e.g., web/mobile applications) seamlessly integrated with Electronic Health Records (EHRs) for actionable insights.
- Global Health Applications: Expanding the model applicability and validation to diverse populations and resource settings worldwide to address the health disparities.

REFERENCES

[1] M. Jalil J. Ghrabat *et al.*, "Utilizing Machine Learning for the Early Detection of Coronary Heart Disease," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 17363–17375, Oct. 2024, https://doi.org/10.48084/etasr.8171.

[2] S. M. Alanazi and G. S. M. Khamis, "Optimizing Machine Learning Classifiers for Enhanced Cardiovascular Disease Prediction," *Engineering, Technology & Applied Science Research*, vol. 14, no. 1, pp. 12911–12917, Feb. 2024, https://doi.org/10.48084/etasr.6684.

[3] S. Saha, Md. M. Rahman, T. T. Suki, Md. M. Alam, Md. S. Alam, and M. A. S. Haque, "Heart Disease Prediction Using Machine Learning Algorithms: Performance Analysis," in *2024 3rd International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, Apr. 2024, pp. 1–6, https://doi.org/10.1109/ICAEEE62219.2024.10561820.

[4] S. P. Singh, A. Singh, and M. Kumari, "Beyond Traditional Methods: Utilizing Regularized Logistic Regression for Accurate Heart Disease Forecasting," in *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)*, May 2024, pp. 247–251, https://doi.org/10.1109/INNOCOMP63224.2024.00048.

[5] P. AmoghVarshith, D. Harika, P. Priyanka, S. S. Kumar, and D. Haritha, "Heart Disease Prediction Using Machine Learning Algorithms," in *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, Apr. 2024, vol. 1, pp. 1–5, https://doi.org/10.1109/ICKECS61492.2024.10616888.

[6] Y. R. Maramreddy and K. Muppavaram, "Detecting and Mitigating Data Poisoning Attacks in Machine Learning: A Weighted Average Approach," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15505–15509, Aug. 2024, https://doi.org/10.48084/etasr.7591.

[7] P. Dhaka, R. Sehrawat, and P. Bhutani, "An Innovative Approach to Cardiovascular Disease Prediction: A Hybrid Deep Learning Model," *Engineering, Technology & Applied Science Research*, vol. 13, no. 6, pp. 12396–12403, Dec. 2023, https://doi.org/10.48084/etasr.6503.

[8] R. Bhuvana, S. Maheshwari, and S. Sasikala, "Predict the Heart Disease Using a Logistic Regression Classifier Algorithm," in *2023 12th International Conference on System Modeling & Advancement in Research Trends (SMART)*, Dec. 2023, pp. 649–652, https://doi.org/10.1109/SMART59791.2023.10428486.

[9] A. K. Dubey, A. K. Sinhal, and R. Sharma, "An Improved Auto Categorical PSO with ML for Heart Disease Prediction," *Engineering, Technology & Applied Science Research*, vol. 12, no. 3, pp. 8567–8573, June 2022, https://doi.org/10.48084/etasr.4854.

[10] S. Usha and S. Kanchana, "Effective Analysis of Heart Disease Prediction using Machine Learning Techniques," in *2022 International Conference on Electronics and Renewable Systems (ICEARS)*, Mar. 2022, pp. 1450–1456, https://doi.org/10.1109/ICEARS53579.2022.9752132.

[11] I. M. El-Hasnony, O. M. Elzeki, A. Alshehri, and H. Salem, "Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction," *Sensors*, vol. 22, no. 3, Jan. 2022, Art. no. 1184, https://doi.org/10.3390/s22031184.

[12] A. Al Bataineh and S. Manacek, "MLP-PSO Hybrid Algorithm for Heart Disease Prediction," *Journal of Personalized Medicine*, vol. 12, no. 8, Aug. 2022, Art. no. 1208, https://doi.org/10.3390/jpm12081208.

[13] N. Mohan, V. Jain, and G. Agrawal, "Heart Disease Prediction Using Supervised Machine Learning Algorithms," in *2021 5th International Conference on Information Systems and Computer Networks (ISCON)*, Oct. 2021, pp. 1–3, https://doi.org/10.1109/ISCON52037.2021.9702314.

[14] Y. Muhammad, M. Tahir, M. Hayat, and K. T. Chong, "Early and accurate detection and diagnosis of heart disease using intelligent computational model," *Scientific Reports*, vol. 10, no. 1, Nov. 2020, Art. no. 19747, https://doi.org/10.1038/s41598-020-76635-9.