International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN (Online): 2347-5552, Volume-12, Issue-6, November 2024
https://doi.org/10.55524/ijircst.2024.12.6.13
Article ID IRP-1574, Pages 95-100
http://www.ijircst.org

A Comparative Study on Predicting Cardiovascular Disease Using Machine Learning Algorithms

Ananya Sarker[1], Md. Harun Or Rashid[2], Arzuman Akhter[3], Ayesha Siddiqua[4], Shafriki Islam Shemul[5], and Must. Asma Yasmin[6]

[1] Assistant Professor, Department of CSE, Bangladesh Army University of Engineering & Technology, Natore, Bangladesh
[2] Lecturer, Department of CSE, Bangabandhu Sheikh Mujibur Rahman University, Kishoreganj, Bangladesh
[3, 4, 5] Alumni, Department of CSE, Bangladesh Army University of Engineering & Technology, Natore, Bangladesh
[6] Associate Professor, Department of CSE, Bangladesh Army University of Engineering & Technology, Natore, Bangladesh

Correspondence should be addressed to Ananya; Sarker; ananya.ruet@gmail.com

Received 29 October 2024; Revised 13 November 2024; Accepted 28 November 2024

ABSTRACT- Heart disease is a global health concern because of eating patterns, office work cultures, and lifestyle changes. A machine learning-based heart attack prediction system is like having a vigilant watchdog in the medical field. To estimate the danger of a heart attack, it all boils down to analyzing data and complex algorithms. Four primary categories were established at the outset of this study: age, gender, BMI, and blood pressure. The data on heart illness was then classified using a variety of machine learning approaches, including XGBoost Model, Gradient Boosting Model, Random Forest, Logistic Regression, and Decision Trees. The results in terms of accuracy, false positive rate, precision, sensitivity, and specificity were then compared. Results in terms of accuracy, precision, recall, and f1_score were found to be greatest when using Logistic Regression (LR). It is therefore strongly recommended that data on cardiac disease can be classified using the logistic regression technique.

KEYWORDS- Heart Disease, Classification, Machine Learning, Precision, Accuracy.

I. INTRODUCTION

Cardiovascular diseases (CVDs), including heart disease, are the leading cause of mortality worldwide, accounting for approximately 17.9 million deaths annually representing nearly one-third of all global deaths [1]. For over 23.6 million people with cardiovascular problems, heart attacks and strokes alone are predicted to be the leading cause of mortality by 2030 [2]. The multifactorial nature of heart disease, involving genetic, environmental, and behavioral risk factors, makes early detection crucial for effective prevention and management. Traditional diagnostic methods, while valuable, often rely on clinician expertise and may not fully utilize the vast and complex patient data generated in modern healthcare systems. Heart attacks and angina (chest pain) are the most frequent type of coronary artery disease (CAD), which is caused by narrowing or blockage of the blood arteries supplying the heart. Other types include heart failure, arrhythmias, heart valve problems, congenital heart defects. The following figure (Figure 1) depicts regarding the variety of issues that affect the heart.

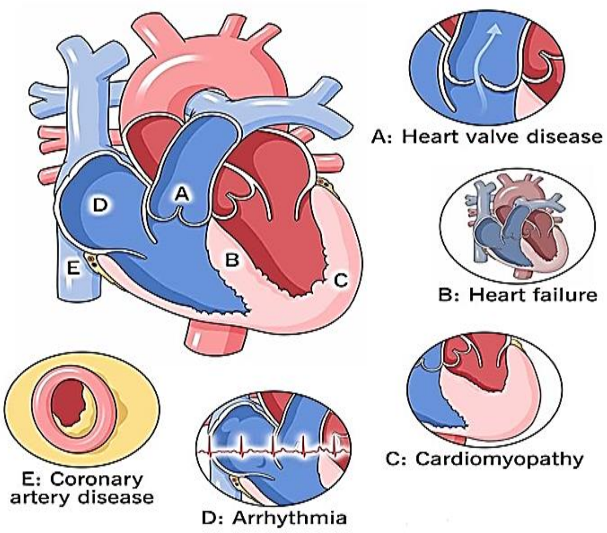


Figure 1: Varieties issue of heart attack [1]

Machine learning (ML), a subset of artificial intelligence, has emerged as a powerful tool for analyzing complex datasets and improving diagnostic accuracy in healthcare. ML algorithms can identify hidden patterns, relationships, and trends in large datasets that are difficult for humans to discern. This capability has made ML particularly useful in predicting heart disease, enabling the development of data-driven models that support clinical decision-making [3].

Various ML techniques, such as logistic regression, decision trees, random forests, support vector machines (SVM), and deep learning models, have been applied to heart disease prediction. These methods vary in terms of accuracy, computational efficiency, and interpretability, which are critical factors for real-world applications [4]. A comparative analysis of these approaches is essential to identify the most effective algorithms for heart disease prediction and to understand their strengths and limitations.

This study aims to provide a comparative study of commonly used machine learning algorithms for predicting heart disease. By evaluating their performance and practical utility, the study seeks to contribute to the growing body of research that leverages artificial intelligence to enhance cardiovascular healthcare outcomes.

II. LITERATURE REVIEW

The threat of heart attacks looms big in our dynamic environment of fast-paced lifestyles and shifting daily routines. The state of our hearts is greatly influenced by our lifestyles, diets, and work environments. However, there is a glimmer of hope, and it comes in the form of technology, particularly machine learning. Think of it as a digital guardian angel when it comes to predicting heart attacks. It delves into our medical data, striving to foresee any potential threats to our hearts. It's not just a prediction; it's a life-saving precaution. This technology has the remarkable ability to uncover hidden patterns and connections that might elude traditional diagnostic methods. Within the field of medicine, where early procedures frequently fail to forecast heart illness, we dive into the world of heart attack prediction in our investigation. The field of heart attack prediction is changing as a result of data analysis and modern technology. With the growing worry over heart disease, machine learning appears to be a potential future direction. We may be able to save lives by using it to anticipate the likelihood of a heart attack well in advance. This is concrete reality that has been impacted by machine learning techniques, not science fiction.

In recent times, several research that use machine learning techniques to predict diabetes have been found in the literature. Sifting through enormous datasets, algorithms like Random Forest, Logistic Regression, and Neural Networks function as watchful sentinels, revealing minute patterns and signals that could indicate an imminent heart attack.

A study was conducted by C. Salkar on Kidney and Heart Disease Prediction using Machine Learning and it was found that Random Forest algorithm obtained an accuracy of 83% [2]. By using multiple machine learning techniques, the authors of another study showed that logistic regression had an accuracy of 75.30% [5].

A. Grag et al. [6] carried out their research on heart disease prediction and found that naïve bayes model performed better with an accuracy 92%. The author in [7] discovered that by concentrating on early detection, the SVM was a good fit for improving accuracy and it was 99.00%. In [8], the authors showed how the most precise prediction was achieved using random forest and the success rate was 86%. The Global Burden of Disease (GBD) 2019 is a global study that assesses disease burden by age, sex, and region, providing crucial insights into health disparities [9]. The authors showed that maximum accuracy was obtained by neural network model and it was 93%. Without these various research had been accomplished on heart disease prediction [10][11][12][13]. Based on the aforementioned research, it can be concluded that a machine learning model is suitable for predicting cardiac disease.

III. METHODOLOGY

In this research, the work started by collecting the heart disease dataset and splitting it into a training set and a testing set. Then preprocessing these data so that these can be fit for feature extraction, and classification techniques are applied to thepreprocessed dataset. Several performance measures were taken into consideration as five different categorization algorithms were implemented and compared with one another. The following figure (Figure 2) depicts the working flow of our approach.

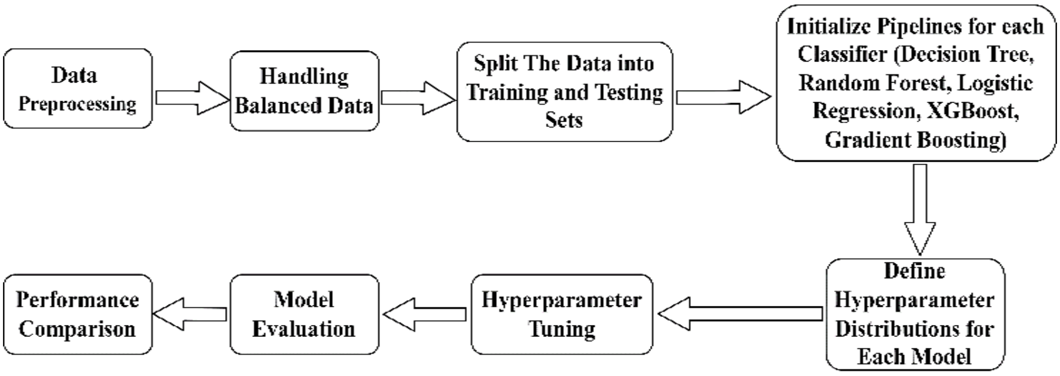

Data Preprocessing
Handling Balanced Data
Split The Data into Training and Testing Sets
Initialize Pipelines for each Classifier (Decision Tree, Random Forest, Logistic Regression, XGBoost, Gradient Boosting)
Define Hyperparameter Distributions for Each Model
Hyperparameter Tuning
Model Evaluation
Performance Comparison

Figure 2: Workflow diagram

The classification techniques applied to the heart disease dataset are acquired from the developer account. The table below (Table 1) shows detailed information regarding the considered dataset.

Table 1: The Description of the Dataset

Criteria	Description
Source	Kaggle
File Type	CSV format
Record	270 tuple, 17 attribute
Classification Type	Age, Gender, BP, Chest Pain, Cholesterol

The following figures (Figure 3, Figure 4, Figure 5 and Figure 6) represent the internal information of the considered dataset.

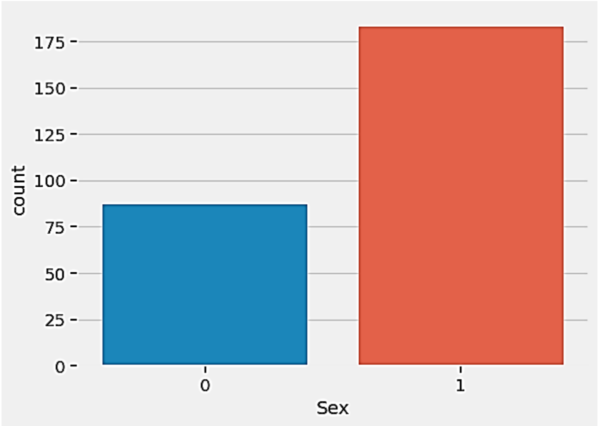


Figure 3: Count of sex (female- 0, male- 1)

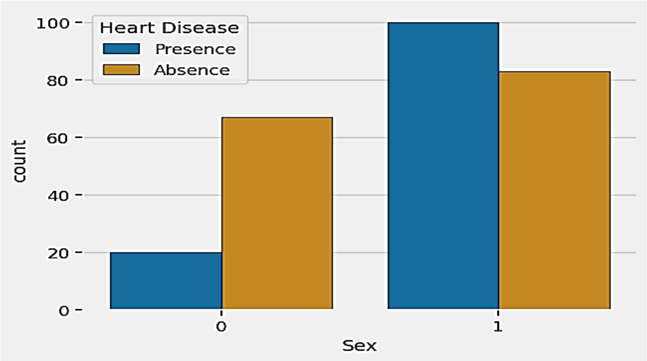
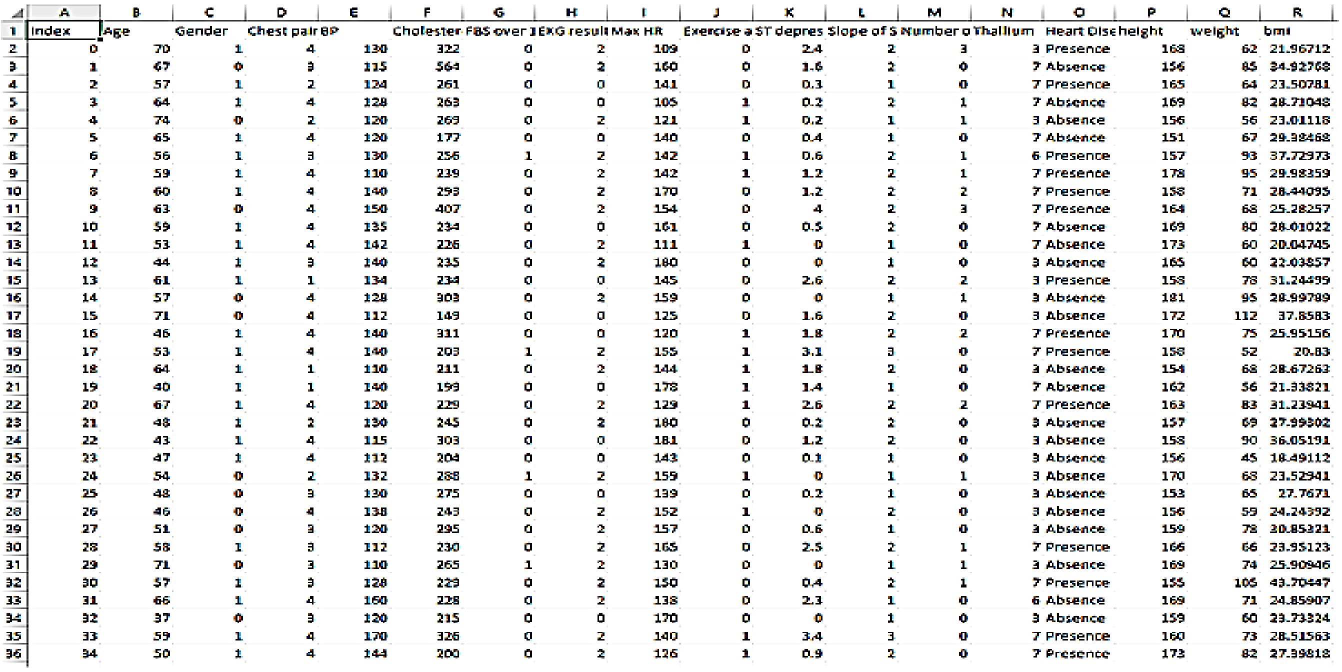


Figure 4: Presence and absence of disease based on Sex



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Index	Age	Gender	Chest pair	BP	Cholester	FBS over 1	EKG result	Max HR	Exercise a	ST depres	Slope of S	Number o	Thallium	Heart Dise	height	weight	bmi
2	0	70	1	4	130	322	0	2	109	0	2.4	2	3	3	Presence	168	62	21.96712
3	1	67	0	3	115	564	0	2	160	0	1.6	2	0	7	Absence	156	85	34.92768
4	2	57	1	2	124	261	0	0	141	0	0.3	1	0	7	Presence	165	64	23.50781
5	3	64	1	4	128	263	0	0	105	1	0.2	2	1	7	Absence	169	82	28.71048
6	4	74	0	2	120	269	0	2	121	1	0.2	1	1	3	Absence	156	56	23.01118
7	5	65	1	4	120	177	0	0	140	0	0.4	1	0	7	Absence	151	67	29.38468
8	6	56	1	3	130	256	1	2	142	1	0.6	2	1	6	Presence	157	93	37.72973
9	7	59	1	4	110	239	0	2	142	1	1.2	2	1	7	Presence	178	95	29.98359
10	8	60	1	4	140	293	0	2	170	0	1.2	2	2	7	Presence	158	71	28.44095
11	9	63	0	4	150	407	0	2	154	0	4	2	3	7	Presence	164	68	25.28257
12	10	59	1	4	135	234	0	0	161	0	0.5	2	0	7	Absence	169	80	28.01022
13	11	53	1	4	142	226	0	2	111	1	0	1	0	7	Absence	173	60	20.04745
14	12	44	1	3	140	235	0	2	180	0	0	1	0	3	Absence	165	60	22.03857
15	13	61	1	1	134	234	0	0	145	0	2.6	2	2	3	Presence	158	78	31.24499
16	14	57	0	4	128	303	0	2	159	0	0	1	1	3	Absence	181	95	28.99789
17	15	71	0	4	112	149	0	0	125	0	1.6	2	0	3	Absence	172	112	37.8583
18	16	46	1	4	140	311	0	0	120	1	1.8	2	2	7	Presence	170	75	25.95156
19	17	53	1	4	140	203	1	2	155	1	3.1	3	0	7	Presence	158	52	20.83
20	18	64	1	1	110	211	0	2	144	1	1.8	2	0	3	Absence	154	68	28.67263
21	19	40	1	1	140	199	0	0	178	1	1.4	1	0	7	Absence	162	56	21.33821
22	20	67	1	4	120	229	0	2	129	1	2.6	2	2	7	Presence	163	83	31.23941
23	21	48	1	2	130	245	0	2	180	0	0.2	2	0	3	Absence	157	69	27.99302
24	22	43	1	4	115	303	0	0	181	0	1.2	2	0	3	Absence	158	90	36.05191
25	23	47	1	4	112	204	0	0	143	0	0.1	1	0	3	Absence	156	45	18.49112
26	24	54	0	2	132	288	1	2	159	1	0	1	1	3	Absence	170	68	23.52941
27	25	48	0	3	130	275	0	0	139	0	0.2	1	0	3	Absence	153	65	27.7671
28	26	46	0	4	138	243	0	2	152	1	0	2	0	3	Absence	156	59	24.24392
29	27	51	0	3	120	295	0	2	157	0	0.6	1	0	3	Absence	159	78	30.85321
30	28	58	1	3	112	230	0	2	165	0	2.5	2	1	7	Presence	166	66	23.95123
31	29	71	0	3	110	265	1	2	130	0	0	1	1	3	Absence	169	74	25.90946
32	30	57	1	3	128	229	0	2	150	0	0.4	2	1	7	Presence	155	105	43.70447
33	31	66	1	4	160	228	0	2	138	0	2.3	1	0	6	Absence	169	71	24.85907
34	32	37	0	3	120	215	0	0	170	0	0	1	0	3	Absence	159	60	23.73324
35	33	59	1	4	170	326	0	2	140	1	3.4	3	0	7	Presence	160	73	28.51563
36	34	50	1	4	144	200	0	2	126	1	0.9	2	0	7	Presence	173	82	27.39818

Figure 5: Internal structure of the dataset

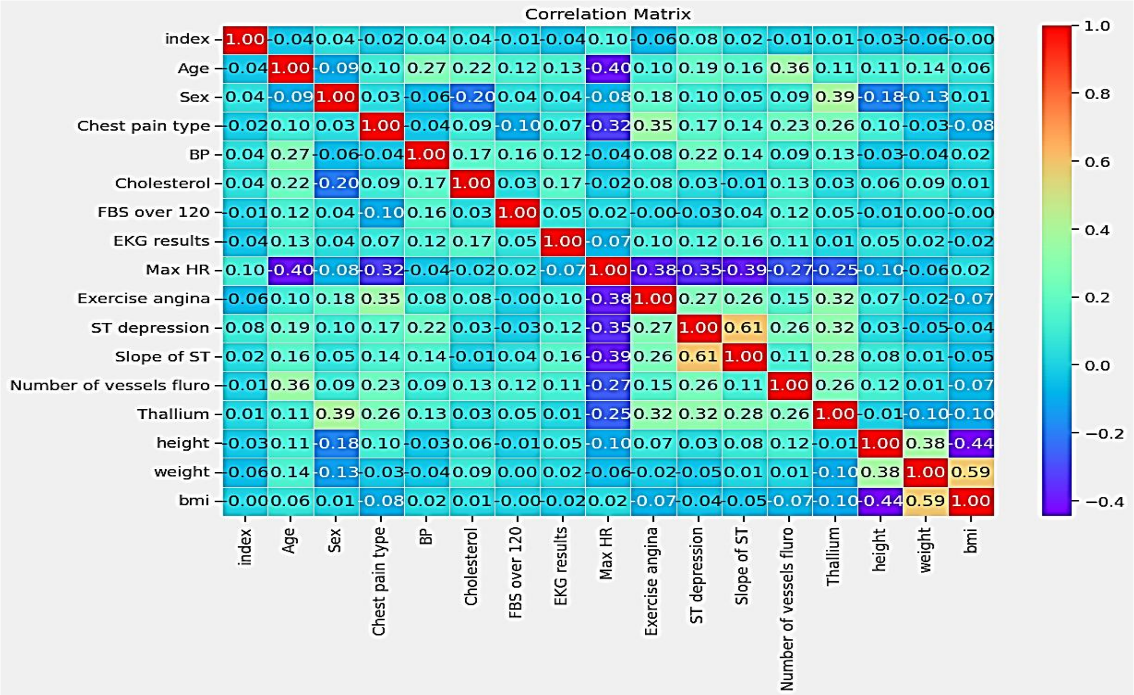


Figure 6: Correlation of numeric features

Among various machine learning techniques Decision Tree, Random Forest, Logistic Regression, XGBoost, and Gradient Boosting techniques are considered in this study.

A. Decision Tree(DT)

Decision tree algorithm is like a roadmap for predicting heart attacks. It works by breaking down the problem into smaller, easier-to-understand steps, just like how we make decisions in our daily lives. It uses data to create a tree-like structure, where each branch represents a choice or a condition related to heart health. For heart attack prediction, these decision trees consider various factors such as age, cholesterol levels, and blood pressure. By following the branches of the tree, the algorithm can pinpoint the likelihood of a heart attack based on a person's specific characteristics. It's kind of like how we make decisions: we assess different aspects of a situation to conclude. This approach is useful because it can reveal the most critical factors that contribute to heart attacks in a way that's easy to interpret. It's not just about predicting heart attacks; it also helps us understand why they might happen. So, decision tree algorithm is a valuable tool in the effort to prevent and manage heart health issues.

B. Random Forest (RF)

Random Forest is a classification algorithm with multiple deep DTs, but can introduce overfitting and make mistakes due to their susceptibility to training results. Think of a random forest algorithm as a team of experts coming together to help you make a decision. In this case, the decision is about the likelihood of someone having a heart attack. Each expert (or tree) has its perspective based on different aspects of a person's health. Imagine you have a panel of doctors, each looking at specific factors like age, cholesterol levels, blood pressure, and more. These experts give their individual opinions, but instead of just relying on one of them, the random forest algorithm combines their insights. The power of this approach is that it takes into account a variety of factors and different viewpoints, just like a group of doctors with diverse expertise. By considering multiple perspectives, the random forest algorithm can make a more accurate prediction about the risk of a heart attack. It's like having a medical team working together to provide you with a comprehensive assessment, ultimately helping in early detection and prevention of heart problems.

C. Logistic Regression (LR)

Logistic Regression is like having a conversation with your heart. It's a method that helps us understand the chances of someone having a heart attack based on various factors. Just like when you talk to a friend and weigh different possibilities, it considers things like age, blood pressure, and cholesterol levels to make a prediction. Imagine it as if we're looking at a group of people and asking, "What's the likelihood of each person having a heart attack?" It takes all these individual possibilities and combines them into a single, easy-to-understand outcome. It's a bit like making a well-informed guess to help us take better care of our hearts.

D. XGBoost

XGBoost is an exciting algorithm when it comes to predicting heart attacks. It's like having a super-smart assistant that uses a blend of decision trees to analyze lots of data and make highly accurate predictions. It's a bit like when a doctor examines all your medical history to tell if you're at risk of a heart attack, but XGBoost does it with lightning speed and incredible precision. It has gained a lot of attention for its ability to spot subtle patterns in the data, helping doctors and researchers make earlier and more reliable predictions about who might be at risk of a heart attack. It's like having an extra set of expert eyes to detect those warning signs that might otherwise go unnoticed. The beauty of XGBoost is that it's not just about crunching numbers; it's about saving lives. By harnessing the power of this algorithm, we're taking significant steps towards preventing heart attacks and ensuring that people get the care they need in time. It's a game-changer in the world of heart disease prediction.

E. Gradient Boosting

With gradient boosting, each new model is trained to minimize the loss function, such as mean squared error or cross-entropy of the preceding model. It is a potent boosting procedure that turns multiple weak learners into strong learners. The approach calculates the gradient of the loss function about the current ensemble's predictions for each iteration and then trains a new weak model to minimize this gradient. Next, the new model's predictions are included in the ensemble, and the procedure is continued until a stopping requirement is satisfied.

IV. COMPARATIVE RESULT ANALYSIS

In our study, among various classifier we had evaluated performance metrics for five classifier and made comparison among them in terms of ROC AUC value, precision, recall, f1-score and accuracy. Out of the five classifiers Logistic Regression performed better. The following table (Table 2) shows a comparison among the applied classification algorithms using various performance metrics.

Table 2: Comparison among Applied Classifiers Using Various Performance Metrics

Model Name	ROC AUC	Precision	Recall	F1-Score	Accuracy
Decision Tree	72%	0.68	0.72	0.70	70.00%
Random Forest	95%	0.84	0.93	0.89	88.33%
Logistic Regression	96%	0.90	0.90	0.90	90.00%
XGBoost	95%	0.88	0.76	0.81	83.33%
Gradient Boosting	93%	0.88	0.79	0.84	85.00%

In our study, classifiers performances are observed considering the attribute 'Heart Disk' and without considering 'Heart Disk' separately. In both cases, Logistic Regression algorithm performs better than others. The below figure (Figure 7) depicts a comparison among the applied classification algorithms.

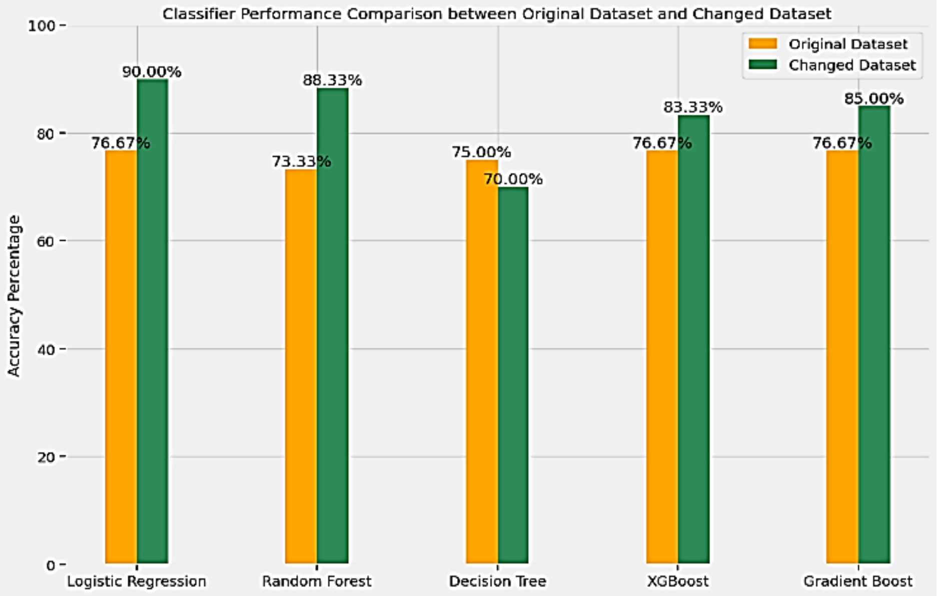


Figure 7: Accuracy comparison with and without heart disc

ROC-AUC is a widely used performance metric for evaluating the effectiveness of classification models, especially in binary classification tasks. It measures the model's ability to distinguish between positive and negative classes. A comparison of the applied classifiers based on the ROC-AUC value is shown in the image below (Figure 8).

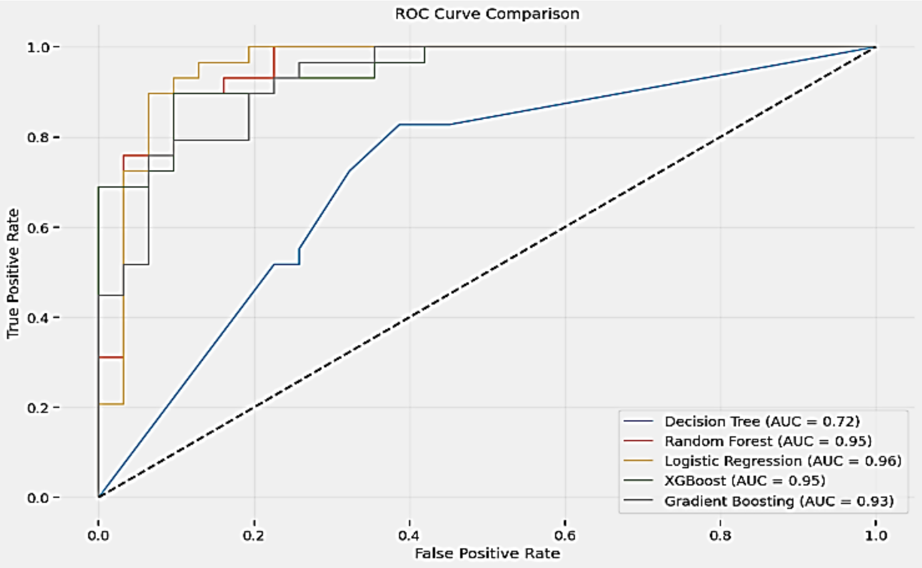


Figure 7: ROC-AUC comparison

V. CONCLUSION

This study explored the use of various machine learning algorithms to predict cardiovascular disease using real-world data. Logistic Regression emerged as the most effective algorithm, valued for its robustness, while other methods like Decision Trees, Random Forest, XGBoost, and Gradient Boosting also showed promise. The study emphasized the need for fine-tuning model parameters and combining methods to enhance accuracy. It concluded by recommending future research focus on larger datasets and advanced machine learning technologies, underscoring the growing role of machine learning in early detection and personalized care for cardiovascular disease.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

ACKNOWLEDGMENT

We like to sincerely show our gratitude to all the authors and organizations who made this research possible.

REFERENCES

[1] World Health Organization, "Cardiovascular diseases (CVDs)," 2021. Available from: https://www.who.int.

[2] C. Salkar, "A detailed analysis on kidney and heart disease prediction using machine learning," *Journal of Computing and Natural Science*, vol. 1, pp. 9-14, 2021. Available from: https://doi.org/10.53759/181X/JCNS202101003

[3] R. Alizadehsani, M. Abdar, M. Roshanzamir, A. Khosravi, P. M. Kebria, F. Khozeimeh, S. Nahavandi, N. Sarrafzadegan, and U. R. Acharya, "Machine learning-based coronary artery disease diagnosis: A comprehensive review," *Computers in Biology and Medicine*, vol. 111, p. 103346, 2019. Available from: https://doi.org/10.1016/j.compbiomed.2019.103346

[4] C. Krittanawong, H. U. H. Virk, S. Bangalore, Z. Wang, K. W. Johnson, R. Pinotti, H. Zhang, S. Kaplin, B. Narasimhan, T. Kitai, U. Baber, J. L. Halperin, and W. H. W. Tang, "Machine learning prediction in cardiovascular diseases: a meta-analysis," *Scientific Reports*, vol. 10, no. 1, p. 16057, 2020. Available from: https://doi.org/10.1038/s41598-020-72685-1

[5] G. Abdulsalam, S. Meshoul, and H. Shaiba, "Explainable heart disease prediction using ensemble-quantum machine learning approach," *Intelligent Automation & Soft Computing*, vol. 36, no. 1, pp. 761-779, 2023. Available from: https://doi.org/10.32604/iasc.2023.032262

[6] A. Garg, B. Sharma, and R. Khan, "Heart disease prediction using machine learning techniques," in *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, p. 012046, 2021. Available from: https://doi.org/10.1088/1757-899X/1022/1/012046

[7] V. Sharma, S. Yadav, and M. Gupta, "Heart disease prediction using machine learning techniques," in *2020 2nd Int. Conf. Adv. Comput., Commun., Control Networking (ICACCCN)*, pp. 177-181, 2020. Available from: https://doi.org/10.1109/ICACCCN51052.2020.9362842

[8] P. Rani, R. Kumar, N. M. O. Sid Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning," *J. Reliab. Intell. Environ.*, vol. 7, no. 3, pp. 263-275, 2021. Available from: https://doi.org/10.1007/s40860-021-00133-6

[9] D. E. Salhi, A. Tari, and M.-T. Kechadi, "Using machine learning for heart disease prediction," in *Adv. Comput. Syst. Appl.*, pp. 70-81, 2021. Available from: https://doi.org/10.1007/978-3-030-69418-0_7

[10] V. Chang, V. R. B. Bhavani, A. Q. Xu, and M. A. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms," *Healthcare Analytics*, vol. 2, p. 100016, 2022. Available from: https://doi.org/10.1016/j.health.2022.100016

[11] 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. Available from: https://doi.org/10.3390/a16020088

[12] V. V. Ramalingam, A. Dandapath, and M. K. Raja, "Heart disease prediction using machine learning techniques: A survey," *Int. J. Eng. Technol.*, vol. 7, no. 2.8, pp. 684-687, 2018. Available from: https://doi.org/10.14419/ijet.v7i2.8.10557

[13] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Computers in Biology and Medicine*, vol. 136, p. 104672, 2021. Available from: https://doi.org/10.1016/j.compbiomed.2021.104672

[14] A. Singh and R. Kumar, "Heart disease prediction using machine learning algorithms," in *2020 Int. Conf. Electr. Electron. Eng. (ICE3)*, pp. 452-457, 2020. Available from: https://doi.org/10.1109/ICE348803.2020.9122958

[15] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in *2008 IEEE/ACS Int. Conf. Comput. Syst. Appl.*, pp. 108-115, 2008. Available from: https://doi.org/10.1109/AICCSA.2008.4493524

[16] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," in *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, p. 012072, 2021. Available from: https://doi.org/10.1088/1757-899X/1022/1/012072

ABOUT THE AUTHORS

Ananya Sarker was born in Natore district of Bangladesh in 1992. She received the Master of Science, M.Sc. Engineering in Computer Science & Engineering (CSE) and Bachelor of Science, B.Sc. Engineering in Computer Science & Engineering (CSE) from the Rajshahi University of Engineering & Technology (RUET), Rajshahi, Bangladesh. Currently, she is pursuing her PhD at the University of Rajshahi. Besides she is working as an assistant professor in the Department of CSE, Bangladesh Army University of Engineering & Technology (BAUET). Her research interests include image processing, machine learning, deep learning and data mining.

Md. Harun Or Rashid was born in Kurigram district of Bangladesh in 1995. He received the Master of Science, M.Sc. Engineering in Computer Science & Engineering (CSE) and the Bachelor of Science, B.Sc. Engineering in Computer Science & Engineering (CSE) from the Rajshahi University of Engineering & Technology (RUET), Rajshahi, Currently, he is working as a Lecturer in the Department of CSE, Bangabandhu Sheikh Mujibur Rahman University, Kishoreganj. His research interests include image processing, machine learning, deep learning and data mining.

Arzuman Akhter was born in Chapainawabganj district of Bangladesh in 2001. She received the Bachelor of Science, B.Sc. in Computer Science & Engineering (CSE) from the Bangladesh Army University of Engineering & Technology (BAUET), Natore, Bangladesh. Currently, she is working as a web developer at a renowned software firm.

Ayesha Siddiqua was born in Natore district of Bangladesh in 2001. She received the Bachelor of Science, B.Sc. in Computer Science & Engineering (CSE) from the Bangladesh Army University of Engineering & Technology (BAUET), Natore, Bangladesh. Currently, she is working as a junior executive engineer at a renowned software firm.

Shafriki Islam Shemul was born in Chapainawabganj district of Bangladesh in 2001. She received the Bachelor of Science, B.Sc. in Computer Science & Engineering (CSE) from the Bangladesh Army University of Engineering & Technology (BAUET), Natore, Bangladesh. Currently, she is working as a front-end developer at a renowned software firm.

Must. Asma Yasmin was born in Hobigonj district of Bangladesh in 1982. Currently, she is pursuing her PhD at the University of Rajshahi. Besides she is working as an associate professor in the Department of CSE, Bangladesh Army University of Engineering & Technology (BAUET). Her research interests include communication engineering and deep learning.